Foundations for Meaning and Understanding in Human-centric AI

Book · June 2022

DOI: 10.5281/zenodo.6666820

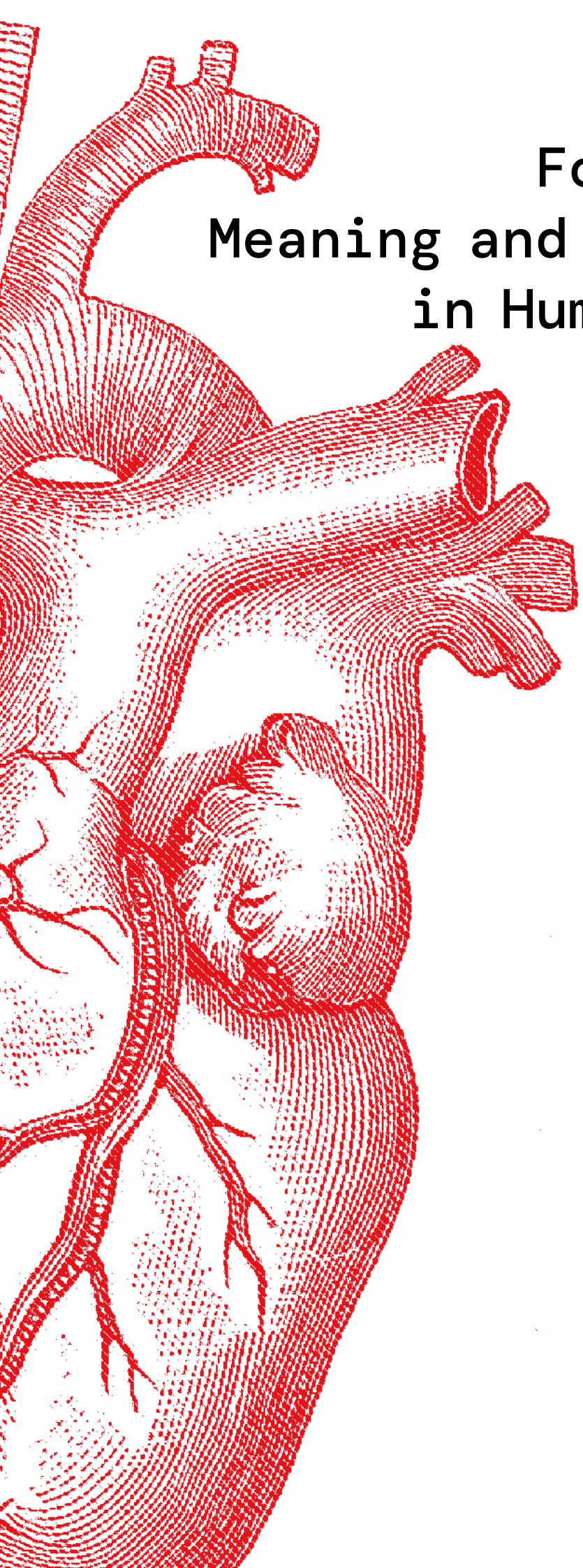
9 authors, including:

Inès Blin
Sony Corporation
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Anna Morbiato
Università Ca' Foscari Venezia
22 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Carlo Romano Marcello Alessandro Santagiustina
Università Ca' Foscari Venezia
34 PUBLICATIONS **58** CITATIONS
SEE PROFILE

Luc Steels
Catalan Institution for Research and Advanced Studies
275 PUBLICATIONS **12,394** CITATIONS
SEE PROFILE



Foundations for Meaning and Understanding in Human-centric AI

Luc Steels editor

MUHAI

Foundations for Meaning and Understanding in Human-centric AI

Editor:
Luc Steels (VIU)

Contributors:

Inès Blin (CSL)
Anna Morbiato (VIU)
Carlo Santagiustina (VIU)
Luc Steels (VIU)
Lise Stork (VUA)
Annette Ten Teije (VUA)
Ilaria Tiddi (VUA)
Remi van Trijp (CSL)
Oscar Vilarroya (IMIM)

Edited by Luc Steels, Venice International University

Cover Graphic Design - studio +fortuna, Trieste

Published by Venice International University, Isola di San Servolo 30133 Venice, Italy

Open Access: https://zenodo.org/communities/muhai/

Citation:

Steels, L. (ed.) Foundations for Meaning and Understanding in Human-centric AI. Venice International University, Venice.

DOI: 10.5281/zenodo.6602456

Volume published in the framework of the MUHAI project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951846. It reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Pathfinder Project under Grant Agreement N° 951846

Executive Summary

The MUHAI consortium studies how it is possible to build AI systems that rest on meaning and understanding. We call this kind of AI **meaningful AI** in contrast to AI that rests exclusively on the use of statistically acquired pattern recognition and pattern completion.

Because meaning and understanding are rather vague and overloaded notions there is no obvious research path to achieve it. The consortium has therefore set up a task early on in the project to explore how understanding is being discussed and treated in other human-centred research fields, more specifically in social brain science, social psychology, linguistics, semiotics, economics, social history and medicine.

Our explorations have yielded a wealth of insights: about understanding in general and the role of narratives in this process, about possible applications of meaningful AI in a diverse set of human-centred fields, and about the technology gaps that need to be plugged to achieve meaningful AI.

This volume summarizes the outcome of our consultations. It has three main parts:
I. A general introduction,
II. A series of chapters reporting on what understanding means in various human-centered research fields other than AI,
III. A short conclusion identifying key research topics for meaning-based human-centric AI.

Our explorations have yielded a wealth of insights: about understanding in general and the role of narratives in this process, about possible applications of meaningful AI in a diverse set of human-centred fields, and about the technology gaps that need to be plugged to achieve meaningful AI.

Contents

.

Chapter 1

Towards Meaningful Human-Centric AI

Luc Steels

Abstract

This chapter focuses on the conceptual foundations of human-centric AI by discussing a number of fundamental questions: What is the nature of meaning and understanding? Why is understanding needed to make AI more human-centric or 'humane'? Why is emulating understanding in artificial systems hard? Why do we need to combine reactive and deliberative intelligence for human-centric AI? What is the role of narratives in understanding? What are some of the open issues for realizing meaningful AI?

Keywords

Meaning, understanding, human-centric AI, meaningful AI, narrative, conceptual foundations of AI.

Parts of this chapter have been published as Steels, L. (2022) Cognitive Foundations of Human-Centric AI. In: Chetouani, M., V. Dignum, P. Lukowicz and C. Sierra (eds) Advanced course on Human-Centered AI. ACAI 2021 Springer Lecture Notes in Artificial Intelligence (LNAI) Post-Proceedings Volume, Tutorial Lecture Series. Springer Verlag, Berlin.

1.1 Introduction

Despite the success in performance of data-driven AI, it also exhibits 'weird' behavior, because it lacks the capacity to understand and exercise deliberative intelligence. You

can consider this weird behavior humorous (Shane 2021) but when it happens with applications that are operating in the real world for real-life social impact, weird AI behavior becomes dangerous and unethical, for example, when an AI system recommends children to execute life-threatening challenges[1] or is making technocratic social decisions that are unfair and cause hardship for those undergoing those decisions.[2]

In reaction to a growing number of such incidents, there have been calls for the development of 'human-centric' or 'humane' AI. As suggested by Nowak, Lukowicz and Horodecki:

> "Human-centric AI focuses on collaborating with humans, enhancing human capabilities, and empowering humans to better achieve their goals." (Nowak, Lukowicz, and Horodecki 2018)

Human-centric AI has become a focal point of current research and development, particularly in Europe, where the EU Commission's strategy on AI and the AI strategies of many EU member states call for AI that shows human agency and oversight, technical robustness and safety, privacy and data governance, transparency, care for diversity, non-discrimination and fairness, focus on societal and environmental well-being, and accountability.(Von der Leyen and al. 2020)

Research in human-centric AI has called for a change in focus compared to the machine-centered AI typified by data-driven statistical machine learning:

- Human-centric AI systems are asked to be aware of the *goals and intentions* of their users and base their own goals and dialog on *meanings* rather than on statistical patterns of past behavior only, even while allowing that statistical patterns can play a very important role, for example for drastically reducing search or carrying out approximate inference.
- Human goals and values should always take precedence. Respect for human autonomy should be built into the system by design, leading to qualities such as *fairness and respect*.
- Human-centric AI systems are asked to be able to explain their reasoning and learning strategies so that the *decisions are understandable by humans*. Only by emphasizing human understandability will human-centric AI achieve proper explainability and transparency.
- Human-centric AI systems should not only learn by observation or theorizing about reality but also by *taking advice* from humans, as suggested in John McCarthy's original 1958 proposal of the Advice Taker.(McCarthy 1958)

[1] In december 2021 the chatbot ALEXA by Amazon recommended a 10 year old to 'plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs'.

[2] In 2020 a scandal known as the 'toeslagenaffaire' (benefit scandal) hit the Dutch political world forcing a fall of the government. Due to excessive zeal of the tax agency controlling the allocation of child benefits and the use of machine learning on social data (which were supposed to be private) many families were pushed into poverty and experienced devastating legal difficulties.

- Human-centric AI should be able to use *natural communication*, i.e. communication primarily based on human language, not only by mimicking language syntax but, more importantly, using the rich semantics of natural languages, augmented with multi-modal communication channels. This is needed to support explainability, and accountability.
- Human-centric AI systems should have the capacity of *self-reflection* which can be achieved by a meta-level architecture that is able to track decision-making and intervene by catching failures and repairing them. By extension, the architecture should support the construction of a theory of mind of other agents, i.e. how they see the world, what their motivations and intentions are, and what knowledge they are using or lacking. Only through this capacity can AI achieve intelligent cooperation and adequate explicability, and learn efficiently through cultural transmission.
- Finally, human-centric AI systems should reflect the *ethical and moral standards* that are also expected from humans or organisations in our society, particularly for supporting tasks that are close to human activity and interest.

All of these objectives point in the direction of meaningful AI, i.e. AI where meaningful distinctions are used to build rich models of problem situations and knowledge domains and where deliberative reasoning complements reactive behavior. The desired properties of human-centric AI are all very difficult to achieve and certainly far beyond the state of the art. They will not appear by decree. Most importantly they require going towards a hybrid form of intelligence that combines reactive and deliberative AI.(Mitchell 2019)

Two arguments have been raised *against* the hypothesis that meaning, understanding and deliberative intelligence are necessary for advancing AI and its applications. The first argument is that big data and statistical training is sufficient to approximate human intelligence for most application areas of (economic) value. The limitations of reactive AI that are increasingly becoming apparent, namely the lack of robustness, a weak capacity to come up with human-understandable explanations and the difficulty to deal with novel situations and outliers, suggest that this argument is not true, or at least not true for many domains of human interest.

But there is a second counter-argument, namely that even though a deliberative form of intelligence would be beneficial, particularly for human-centric AI, it is an impossible target because the required knowledge is not available and cannot be acquired by machines. Proponents of this argument point to earlier research in AI on understanding and deliberation in the 1970s and 1980s. Although this work lead to an earlier strong wave of AI applications (namely expert systems in the 1990s and the semantic web in the 2000s) they argue this research has stalled once it was realized that understanding requires a massive amount of encyclopedic knowledge, fine-grained accurate language processing based on in-depth grammars, and the ability to categorize perceived reality onto the ontologies required for rich models.

However, the situation has changed compared to decades ago. Not only are there now

more powerful symbolic learning strategies, we also have very large knowledge bases in the form of knowledge graphs that provide some of the needed encyclopedic background knowledge.(Antoniou and Harmelen 2008) They have been made possible by the large-scale collective efforts to feed information to various open encyclopedic resources such as Wikipedia and by AI-based 'crawlers' that scavange the web for information. These knowledge graphs are still expanding at a rapid pace.

Moreover thanks to advances in fine-grained precision language processing, particularly for computational construction grammar, we can now go significantly beyond the coarse-grained surface analysis that is common in statistical NLP and analyze sentences from a semantic and pragmatic point of view.(L. Steels 2012a) In addition, techniques for grounding situation models in real world sensory-motor data have made significant advances as well thanks to important advances in machine vision and dynamic motor control, thanks to data-driven approach.(L. Steels and Hild 2012)

Summarizing:

> **Human-centric AI systems are asked to be aware of the goals and intentions of their users, exhibit fairness and respect, explain their decisions in human terms, take human advice, use natural communication, can self-reflect and follow ethical and moral standards. All this requires going beyond data-driven reactive AI and integrating the capacity to understand in a human-like narrative way, be sensitive to context, and perform deliberative reasoning in addition to reacting quickly with stimulus-response associations.**

The rest of this paper unpicks the various concepts used in this summary. I will address the following questions: What is the distinction between reactive and deliberative intelligence? What is understanding? What are meanings? Why is understanding hard? And what is the role of narratives in understanding?

1.2 What is the distinction between reactive and deliberative intelligence?

Try to complete the sentence:

(1) Frank cannot come because his wife has tested positive for **...**

Most people (and also search engines) would not hesitate for one second and complete this sentence automatically and quickly with the word 'covid' or a synonym of that word. Kahneman (D. Kahneman 2011) categorises this as a *reactive* form of intelligence, which he calls system 1. It is *fast thinking* - if we could even call this thinking. It is automatic, effortless and without awareness. A fast response is possible when there is an associative memory that directly relates stimuli with responses and these stimulus-response patterns

can be acquired by sufficient exposure to examples and an induction algorithm. In this case, the fast response is possible because the word 'covid' has been appearing a lot in this specific textual context (n-grams) and we (or an AI system) have acquired the statistical frequency of the n-gram 'tested positive for covid'. A decade ago the word 'AIDS' or 'Ebola' would have been more frequent in this n-gram.

Whereas reactive intelligence is found in all animals (and some would argue also in plants), human intelligence is special because it can operate fluently not only in a reactive but also in a deliberative mode (Kahneman's system 2). A *deliberative* mode of intelligence is classified as *slow thinking* because it takes more time and more effort. A deliberative mode is based on making a rich model of the situation that enables the formation and consideration of different hypotheses. The model should be grounded in the facts known about the situation but also in previous experiences of similar situations. In a deliberative mode we become aware of our reasoning, can verbalize the argumentation, and explain to others why we are thinking in a particular way. We learn in a constructivist fashion by creatively generating, testing and adapting theories and by cultural transmission through language and education.

Using a deliberative mode we can also complete the sentence above but now based on a model that contains not only the stated facts (for example that Frank has a wife and that she tested positive) but also inferences based on common sense knowledge (for example that a husband and a wife typically live together and hence have a lot of exposure to each other's infections) as well as specific knowledge (for example about what rules hold during the covid pandemic in a particular country).

Here are some more facts a rich model for sentence (1) could include:

- Frank was supposed to come to a meeting.
- Covid is an infectious disease
- There is a covid pandemic.
- Covid can be tested.
- A person infected has to stay in quarantine.
- Quarantine means you have to stay home.
- If you have been in contact with an infected person you have to go in quarantine yourself.
- A husband and wife typically live together and hence have a lot of exposure to each other.
- Frank's wife is a high risk contact for Frank.

Given these facts and additional world knowledge, it is possible to answer questions like 'If only Frank's wife has tested positive, why is he himself not able to come to the meeting?' 'Does Frank also have to be tested?', and counterfactual questions like 'Suppose that the meeting alluded to is in Australia and Frank has been spending the last month there without his wife who stayed in the Netherlands, would Frank still not be allowed to come?' The model also supports the formulation of specific hypotheses, such as 'Frank's wife was probably tested recently for Covid' or general hypotheses, such as 'Where Frank lives,

people who have had contact with a positive person are supposed to receive a message to quarantine.'

A rich model gives *insight* into the problem situation and shows paths to problem solutions. It not only helps in question-answering, problem solving or decision making. It prevents inappropriate predictions. Consider for example sentence (2):

(2) Soon after her marriage, Lise was even more happy because she tested positive for **...**

Completion with the word 'pregnancy' is now more appropriate despite the higher frequency of the n-gram 'tested positive for Covid'. Understanding this sentence and providing the most sensible completion requires knowing and inferring facts such as: marriage is typically the beginning of forming a family, getting children requires getting pregnant, pregnancy can be tested, a positive test for pregnancy makes you happy if you want children, testing positive for Covid does not make you happy.

Summarizing:

> **Reactive intelligence rests on ready-made associations between stimulus patterns and responses. Deliberative intelligence rests on making rich models and using reasoning and inference to find a solution. Humans use both forms of intelligence and so should artificial intelligence.**

1.3 Are reactive and deliberative intelligence both needed for AI?

Recent advances in AI have shown that neural networks trained on sufficient amounts of data are able to emulate reactive intelligence and reach unexpected high levels of performance. The training is typically carried out through a prediction task. For example, a large corpus of language texts is assembled and the network is trained to predict the next word in a sentence in this corpus,(Devlin et al. 2019) or a large set of images is assembled and labeled and the network is trained to predict which label should be assigned to an image.(Redmon et al. 2016) Due to the capacity of neural networks to generalize, the prediction is not just based on a simple recall but can handle variations in the input, as long as they stay within the statistical distribution found in the training data.

We call AI systems that emulate reactive intelligence *reactive AI*. Reactive AI is not the exclusive province of neural networks. The earlier rule-based expert systems from the 1970s and 80s mostly tried to solve expert problems by recognizing patterns and finding solutions as fast as possible by the application of heuristics.(Feigenbaum 1977) A lot of the work in behavior-based robotics in the early 1990s also attempted to come to grips with reactive intelligence, but now by the construction of dynamical systems that establish quite direct connections between sensing and actuating.(L. Steels and Brooks 1994)

On the other hand research in deliberative intelligence was the focal point of earlier AI research that emphasized the construction of symbolic (i.e. non-numerical) models, logical inference procedures, the design and implementation of ontologies, the gathering and implementation of large amounts of domain knowledge, symbolic and constructivist learning strategies, and fine-grained precision language processing based on grammars. This research has also lead to impressive demonstrations ranging from mathematical theorem provers to expert systems supporting the configuration of complex equipment.

However, it has also become clear that there is no 'quick fix'. Currently available techniques for artificial deliberative intelligence or *deliberative AI*, such as answer-set programming, constraint solving or logic programming, are powerful but they require that a problem is first exhaustively stated in a logical form and that all needed knowledge is expressed in carefully formalized axioms. Only in that case can the inference machine go to work. It is precisely this first step, namely to define, formalize and structure the problem and the relevant world knowledge, that requires the process we call understanding. Moreover considerable ingenuity and effort is needed to sufficiently catch up and keep up with human competence, which makes deliberative AI not economically viable in many circumstances. But of course that does not mean that deliberative intelligence is not an integral component of human intelligence nor that deliberative AI is crucial in a wide range of applications.

The Achilles' heel of implementing reactive AI using statistical induction is the need for a large amount of representative data, which often is not available or not available in a clean enough form to be reliable. The Achilles' heel of deliberative AI is the need for large amounts of knowledge, which may not exist or cannot be verbalized by human experts. The cooperation between the two modes of intelligence can potentially overcome both limitations and lead us to the next wave of AI. Already many reactive AI systems are trained using the outcome of deliberative intelligence and they fall back on deliberative intelligence when there are no ready-made solution patterns available yet. This approach was first demonstrated by Arthur Samuel in 1959 (Samuel 1959) for the game of checkers but underlies also the success of AlphaGo and other AI systems for board games. At the same time many systems using deliberative intelligence rely on reactive intelligence to tackle larger scale problems, dampen search spaces, provide fast access to very large knowledge bases or deal with grounding issues. So it is a safe bet that the future lies in a combination of reactive and deliberative AI, even for sensory-motor intelligence. But the key lies in mastering the understanding process.

Summarizing:

> **Human intelligence relies both on reactive and deliberative intelligence and so should AI. Reactive intelligence can bring fast solutions to subproblems and agents can therefore solve more challenging problems. On the other hand deliberative intelligence can solve problems where no direct solution is available by exploiting domain knowledge and problems where solutions to subproblems have to be combined in novel ways, but a critical prerequisite is the con-**

struction of rich models from which the relevant information can be derived.

1.4 What is understanding?

Where do the rich models needed for deliberative intelligence come from? They are based on language input, prior knowledge, visual observation, memory of past episodes, mental simulation, inference. The process of constructing a model that integrates all these different sources of information in the service of making the rich models needed for deliberative intelligence is what we call *understanding*. It requires reactive intelligence: for grounding the model in sensory data through feature detection and pattern recognition, for the fast access to possibly relevant information and for the acquisition and application of heuristics to decide which hypothesis deserves to be prioritized. The construction of a rich model also requires deliberative intelligence to fill in gaps through inference and world knowledge.

Here is an example. Consider the image in Fig. 1.1 (left) (adapted from (L. Steels 2020)). This is from a poster that used to be employed in French and Belgian schools to teach children about daily life and to learn how to talk about it. We instantly recognize that this is a scene from a restaurant, using cues like the dress and activities of the waiter and waitress, or the fact that people are sitting at different tables in the room. Data-driven image recognition algorithms are able to segment and identify some of the people and objects in the scene and in some cases label them with a fair degree of accuracy, see Fig. 1.1 (right).

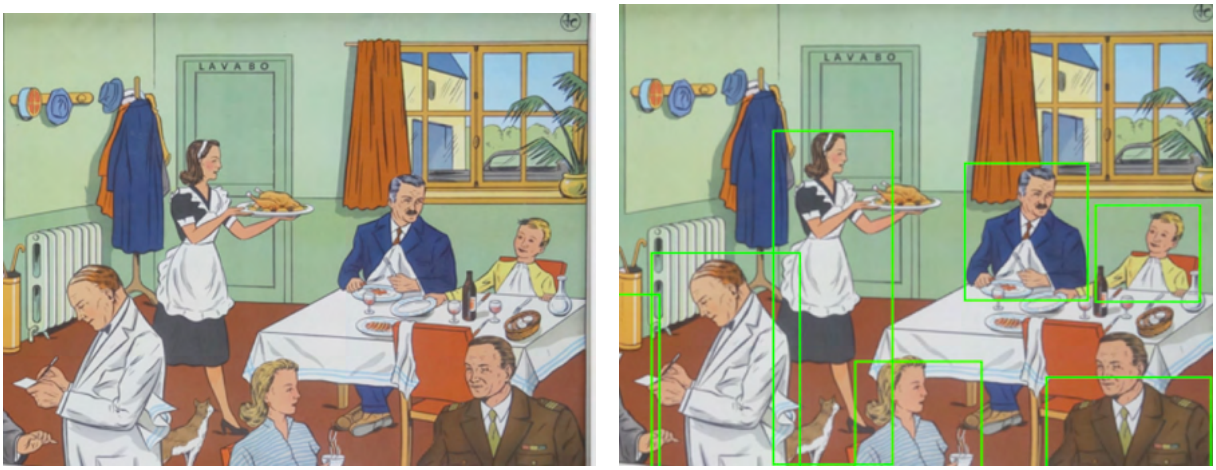


Figure 1.1: Left: Didactic image of a scene in a restaurant. Right: Image segmentation identifying regions that contain people (based on Google's Cloud Vision API).

But these algorithms do not understand the picture in the way we would commonly use this term. Understanding requires a lot more than segmentation and labeling. For example, when asked whether a person is missing at the table on the right, we could all

come up with a straightforward answer: Yes there is a person missing, because there is an empty chair, a plate and cutlery on the table section in front of the chair, and a napkin hanging over the chair. So there must have been a third person sitting there, probably the mother of the child also sitting at the table. Moreover nobody has a lot of difficulty to imagine where the mother went. There is a door marked 'lavabo' (meaning 'toilet' in French) and it is quite plausible that she went to the toilet while waiting for the meal to arrive. Any human viewer would furthermore guess without hesitation why the child is showing his plate to the waitress arriving with the food and why the person to the left of the child (from our perspective) is probably the father looking contently at the child. We could go on further completing the description, for example, ask why the cat at the feet of the waitress looks eagerly at the food, observe that the food contains chicken with potatoes, notice that it looks windy outside, that the vegetation suggests some place in the south of France, and so on.

Clearly these interpretations rely heavily on inferences reflecting knowledge about restaurants, families, needs and desires, roles played by people in restaurants (waiter, waitress, bar tender, cashier, customer). These inferences are not only necessary to properly interpret the visual image in Fig. 1 but also to answer questions such as 'Who is the waitress?', 'Why is she approaching the table?', 'Where is the missing person at the table?, 'Who will get food first?', We can also make predictions and reconstructions, for example, that the waitress will reach the table, put the food on the table, cut the chicken into pieces, and put them on the different plates, or that the mother of the child will come back from the toilet, sit down again at the table, and start eating herself.

Summarizing:

> **Understanding is the process of constructing a meaningful model of a situation and linking this model to background knowledge, memory of similar past situations and factual observations. The model needs to be sufficiently detailed to support deliberative intelligence in tasks like answering questions, giving explanations for these answers, generating hypotheses, seeking more evidence, handling counterfactuals, inferring a plan of action and making predictions in a deliberative way.**

1.5 What are meanings?

The definition of understanding given earlier emphasizes that a model must be 'meaningful', but what does that mean? The concept of meaning is notoriously difficult to define and it has many facets. For the present purpose, let us adopt a definition articulated by philosophers Sperber and Wilson (Sperber and Wilson 1969): A model consists of descriptions and the building blocks of these descriptions are *distinctions*, also commonly called categorisations or concepts. We say that **a model is meaningful if the distinctions it uses are relevant to the interaction between the agent making the model and**

the environment, in other words if the descriptions making up the model are critical for the tasks and contexts the agent has to cope with.

For example, the distinction between red and green is meaningful in the context of traffic because red means you should stop and green means you can advance. The exact choice for which region in the spectrum represents red or green is not critical - and its perception would vary with the ambient light and the context anyway. The spectral value is even culturally determined so that in some countries the green traffic light looks rather blue[3] or the orange traffic light amber or yellow. This example makes the point that a perceptual category exploits a regularity in the environment but is not entirely dependent on it and consequently human perceptual categories are not derivable from statistical induction over examples only.

Here is a second example: The distinction between adult and non-adult is meaningful in the context of law because it refers to the moment at which one acquires full capacities and responsibilities under the law. The age is usually defined as 18 years, but that can differ between countries (in Scotland it is 16 years) and purposes (like obtaining a driver's licence which may be allowed earlier). Moreover, in many countries (like Belgium) you can also be classified as adult (in the sense of gaining full legal capacity) if you are a parent or have gotten married *before* 18 years. The meaning of adult is therefore rooted in interactions between the agent and the environment, which now includes other agents and legal institutions and customs. This example makes again the point that human distinctions cannot be learned empirically from observations alone. As Wittgenstein said: Meaningful distinctions are imposed on reality, shared through conventions and procedures, and defined and transmitted through language.

It is not only that the distinctions used in meaningful models must be relevant, also the way the context is segmented into different entities must be relevant. Which entities are included in a model depends on the task and the context for which the model is intended. For example, a sequence of events may be considered as a single episode to be described as a whole or to different episodes, depending on the purposes of the model and the available data. Similarly, how we decompose an object into its subcomponents depends on whether we have to recognize the object or to dismantle and repair it.

Although data-driven statistical AI applications are responsible for the recent headlines in AI, there are indications that the distinctions these applications acquire and use are not meaningful in a human sense. This leads to a lack of robustness and difficulty to provide human-understandable explanations.

Consider image recognition. Although on benchmarks like digit recognition very high scores are reached, adversarial examples show that for *natural* image recognition, segmen-

[3]For example, in Japan the green light looks more blueish because until a century ago Japanese did not have a basic color word for green, only for blue ("ao") and green was considered a shade of ao. Contemporary Japanese has a word for green, "midori", but the traffic light is still called "ao". As a compromise to abide by international regulations but not deviate too much from language custom, traffic lights in Japan are a greenish shade of blue rather than prototypical green.

tation and labeling, there is an important lack of robustness (one pixel can be enough to derail the recognition process), gross misclassification when an object is not exactly in a setting similar to that used in the training set or the object is slightly altered or its orientation shifted.(Szegedy et al. 2014) These limitations show that the acquired statistical models are sensitive to image features and categorisations which are to some extent adequate for a prediction task but only very indirectly related to the way humans visually identify and categorize objects for other tasks - including tasks which require recognizing objects in a way human intelligence would find natural.

Or consider language processing. BERT, a state-of-the-art statistical language model trained with a 300Mi word corpus and a neural bi-directional encoder transformer algorithm, correctly completes "A robin is a ..." with the word "bird", based on statistical frequency, but if you ask for the completion of "A robin is not a ..." it also answers "bird" for the same reason.(Riedl 2020) A human observer cannot help but conclude that BERT does not understand what the sentence is about and is therefore unreliable. BERT has been trained for language prediction but that task is not the same as that of language understanding or language production in communication tasks, which is after all what language is for. This observed lack of robustness in performance will not improve with more data or with a better induction algorithm because it is due to fundamental epistemological limitations of empirically derived knowledge(Pearl and Mackenzie 2019) and the fact that often the analyses which have to be learned are not observable. For example, we cannot expect that the semantic aspects of language grammars can be learned purely based on data of possible language forms.(Bender and A. 2020) To claim that BERT, or similar systems such as GPT-3 which are trained on sentence completion, understand language is confusing the issues and creates unrealistic expectations for the general public.

Coming up with explanations based on statistical models trained in prediction tasks has proven to be a headache as well. There is a large amount of research going on at the moment and many approaches are being tried:(Mohseni, Zarei, and Ragan 2021) Translating the decision making by a statistical black box model into another more comprehensive but still numerical model (for example decision trees); illustrating which features have been used in decision making by highlighting the areas of the input (for example in an image) that played a role; etc. This is all helpful but is only a stopgap to avoid constructing the kind of meaningful explanations that humans are able to provide and expect, i.e. explanations that explicate the background knowledge and the reasoning steps how a conclusion was reached in terms of rich models composed in human-understandable concepts, in other words using distinctions that overlap with those humans find relevant to their own experiences.(Moore and Swartout. 1988)

Summarizing:

> **We say that a model is meaningful if the entities described in the model and the distinctions (categories) and entities being used to form descriptions are relevant to the interaction between the agent and the environment in a specific set of tasks and domains. We call AI based on meaningful models Meaningful AI.**

1.6 Why is understanding hard?

Many problem situations in which humans find themselves have properties that make it very hard to solve them with reactive intelligence alone. But these properties also make it hard to construct the models needed for deliberative intelligence, even though that is the only way to find and justify solutions in many cases. These properties include: indeterminacy, ambiguity, uncertainty, combinatorial complexity and the characteristics of open worlds.

- *Underspecification* arises when the situation or the problem definition does not contain enough information to find a solution. The missing information must be inferred based on prior knowledge.

- *Incompleteness* arises when there is so much knowledge involved in a domain that it cannot all be formalized or foreseen. This is known as the frame problem in AI.(McCarthy and Hayes 1969) The default cases can be described but all the exceptions cannot.

- *Ambiguity* arises when the same signal or sign can have many possible interpretations. The interpretation valid in the specific problem context must then be inferred by examining the consequences of the different possibilities. This is for example the case in parsing and interpreting human language where every word or phrase typically has more than one meaning or function.

- *Uncertainty* arises when facts can only be established approximately, for example because they are based on a measurement that gives only an approximation to what is to be measured or because facts were derived through induction and the data was incomplete or polluted. Medical diagnosis is a typical example of a domain where uncertainty is rampant because not everything can be measured accurately and detailed information of the behavior of a disease is often not available.

- *Combinatorial complexity* arises as a consequence of underspecification, ambiguity, incompleteness and uncertainty and when solution steps need to be chained. In those cases you need to construct a search space to explore different hypotheses and this space can grow exponentially. Board games, such as chess, are the prototypical example of combinatorial complexity, but this issue is equally present in virtually all applications of AI, for example in syntactic and semantic analyis of natural language sentences or in the design of systems where many different components have to be chosen and put together to achieve a particular function.

- *Open worlds* arise when problem situations come up that are significantly different from anything seen before, i.e. not only variations within known statistical boundaries or new facts that deviate from templates derived from earlier facts. This issue is the norm for socio-economic applications or applications which involve living systems. It is also the case for handling natural language where novel sounds, words,

phrases, constructions and pragmatic interaction patterns may appear, sometimes at a rapid pace.

Certainly, there are domains where the world is closed and the basic rules do not change. There are no ambiguous inputs, full knowledge is available of the situation, and the combinatorics can be worked out and computed in advance or sufficiently tamed with powerful heuristics to guide the search. A board game like checkers and most other adversarial games are examples of this. The rules of checkers are fixed, the board positions are totally determined and observable without ambiguity and the search space is large but computable. Recent progress in game playing (for chess, Go, and other games) is due to massive increases in computer power so that many more games can be tried and heuristics for navigating much bigger search spaces can be learned than before.

On the other hand there are plenty of domains of high human interest where these issues are very much present. These are the domains that cannot be solved in an algorithmic way and have therefore been traditionally the core research target of AI. They require that the AI system understands.

All the above issues were identified from the very beginning of the field in the 1950s, for example, in the work of Herbert Simon on bounded rationality(Simon 1969) or the work of John McCarthy on non-monotic reasoning to cope with incompleteness and open worlds.(McCarthy 1958) To illustrate the issues further, I briefly introduce an example from the domain of common sense problem solving which has received considerable attention in recent AI research, namely cooking recipes.(Beetz et al. 2012)

Nothing seems to be more straightforward than preparing a dish from a recipe, except when you start cooking, and then it often turns out to be hard. I am not referring to all the skill required to handle food but of reading and understanding the recipe, in other words making a sufficiently rich model of the food preparation process so that concrete cooking actions can be undertaken. Part of the problem is to figure out what items and actions the descriptions in the recipe refer to, which objects are involved in each action, and what the parameters of actions are, such as how long something needs to be cooked. Often you also need to come up with alternative solutions to reach the same (implicit) goals.

Consider for example a recipe for preparing *Linguine con lenticchie e pancetta* (Linguine pasta with lentils and pancetta)[4]. The recipe starts with a list of the ingredients and some initial preparations:

> 200 g. small brown lentils from Umbria.
> Two tablespoons of extra virgin olive oil.
> 125g. Pancetta cut into small strips.
> 400g of linguine, etc.

The first challenge is already to find these ingredients or decide on alternatives depending on their availability in your kitchen or in local shops. Although Umbria is renowned for lentils, other regions, like Puy in the Auvergne in France, produce excellent lentils as

[4]Werle, L. (2009) La Cucina della Mamma. Allegrio, Olen (BE), p. 22

well. The pancetta can be replaced by bacon, the linguine by pappardelle. Notice that generating and selecting these alternatives already rests on significant knowledge of the different types of cooking ingredients and their functions in a recipe. For example, the pancetta can be replaced by bacon because they are both salt-cured pork belly salume, the linguine by pappardelle because they are both long dried pasta.

The first line of the recipe says: 'Place the lentils in a pan and add enough water to cover them with a layer of 5 centimeters.' Some of the words in this sentence are ambiguous. For example 'pan' means here a cooking pan but the word can also refer to the action of taking a wider view with a camera. 'Cover' means here 'to put something over something else' but it can also mean 'the object with which you cover something', 'to deal with a subject', 'the outside of a book', 'a place to go hiding'. We have such powerful context-sensitive parsing mechanisms that we do not even notice these ambiguities.

Then there is underspecification. The recipe talks about 'a pan' assuming there is such an object available in the context. 'Them' is presumably refering to the lentils. The water has to be added but to what - presumably to the pan. The recipe does not specify the exact quantity of water to be added but describes instead the end state ('until there is a layer of 5 cm'.).

The next line says 'Add salt and put the lid on the pan.' The salt is not mentioned in the ingredients but assumed to be present by default. Neither does the recipe specify what kind of salt should be added; regular kitchen salt, sea salt, Himalayan black salt? 'Add salt', yes - but to what? How much salt? 'The lid', but which lid? The lid has not been mentioned yet but because a pan has been mentioned and pans have lids, we know that it must be the lid of a pan. But which pan? Presumably the one mentioned in the first line of the recipe.

Clearly, a recipe is not like a computer program which is fully determinate and unambiguous. Understanding the recipe so that you can cook the dish, requires disambiguating, contextualizing and grounding the objects mentioned in order to resolve ambiguities and handle underspecification and incompleteness as much as needed. Doing this requires a significant amount of background knowledge but also the ability to build models of reality based on sensori-motor input, perform mental simulation of actions, and link the referents of descriptions in the recipe to the model.

The cooking domain is representative of a much larger set of domains with similar properties such as instructions for putting together IKEA furniture, manuals for installing and operating equipment, office procedures, legal procedures, instructions for a field trip in the forest. And there are many more domains where the same issues arise. Here are two more examples: making sense of events and medical or technical diagnosis.

All of us are constantly trying to understand the connections between events, both at a personal level and at a historical and political level, and there are scientific fields such as history, criminology, anthropology or journalism that are specialized in trying to piece together what has happened from bits and pieces of indeterminate, ambiguous and uncertain information and structure them into a coherent whole. This is another class of task

challenges where the construction of rich models is central even though the 'true' sequence of events is often impossible to reconstruct, particularly for turbulent historical events, such as a coup d'état or a revolution. The efforts at the moment by a special parliamentary commission to document and understand the attack on the United States Capitol on January 6, 2021 shows how difficult the exercise is, even for recent events. In the case of events further in the past this reconstruction often has to happen from accounts that are only partly true due to lack of knowledge or because the available accounts warp the facts to convey particular viewpoints to serve certain agendas.

Here is another example: medical or technical diagnosis. Both rest on a set of inputs that need to be puzzled together into a coherent model of a disease or a malfunction, which is sufficiently detailed to allow treatment or repair. The inputs are almost always incomplete. It may be possible to perform additional measurements but these measurements typically still yield uncertain outcomes and they may be too expensive or risky to perform. Knowledge about many diseases or causes of faults is often incomplete and continously evolving, not only because knowledge advances thanks to science but also because diseases themselves may change as the biological organisms that cause them (viruses and bacteria) change. The Covid pandemic is a contemporary example, where the medical profession as well as politicians and the public are visibly trying to cope with the changing nature of a virus and the unpredictable effect of rules to contain the pandemic. Here we see the process of understanding in action. It is another example domain showing not only why understanding is central to human intelligence, but also why it is hard.

Summarizing:

> **Understanding is hard in many tasks of human interest because of the indeterminacy, ambiguity and uncertainty of inputs and available knowledge. In addition, there may be rampant combinatorial complexity if different hypotheses have to be explored, and we may need to cope with novel situations due to the fact that the real world is open and in constant flux.**

The issues listed here are experienced by humans and by machines alike. But human intelligence has distinct advantages: We collect throughout life a massive amount information that is relevant for our interactions with the world and others. We have all the necessary subsystems ranging from sensory-motor intelligence, language processing, mental simulation, semantic and episodic memory and learning to meta-level reflection and affective response generation 'under one roof' and we seamlessly integrate all these capacities in the service of solving hard problems. All of this is truely astounding and an enormous challenge to emulate in artificial systems. The argument, sometimes heard these days, that 'Artificial General Intelligence' that will surpass human intelligence is just around the corner (Tegmark 2017), underestimates human intelligence and/or overestimates the true state of the art in AI. Mastering understanding is the hard problem of AI and achieving it will require considerable breakthroughs and concerted effort.

1.7 What are digital twins?

Let us now turn to the kind of models that are required for deliberative intelligence. I mentioned earlier that they have to be 'rich'. What should we mean by that? As a starting point, we need to make a distinction between models that are about the real world and physical interaction in that world and models that primarily attempt to interpret the world and the intentions, motivations and goals of actors in the world.

Real world models should reflect reality as close as possible. They are known as *digital twins*, maintaining a computational representation based on perception of reality and using inferences that reflect how the world changes when certain actions are carried out.(Kuempel, Mueller, and Beetz 2021) A cooking robot needs such a digital twin about the kitchen, representing the kitchen state, possible actions, and the recipe being cooked. Medical diagnosis and treatment also relies on a digital twin, this time about the patient. It represents the state of a patient, hypotheses about possible disease factors and their effect on the body and how therapy may heal the patient. Given the infinite complexity of the real world a digital twin does not have the same level of detail as the actual object it is modelling. It just have to have enough detail for the purposes of the model.

Interpretive models are called *narratives* in humanistic and social sciences. Narrative models not only represent the facts about a situation but also a perspective, framing and conceptualization of the facts (called a plot) and a way of structuring the facts and the plot in terms of semiotic objects such as a textual stories or figurative images (called a narration). Narrative models are discussed in the next section.

Real world models are rich because they typically have multiple layers:

- *Perceptual models* are directly grounded in real world perceptual or proprioceptive data in a robot or in measurement devices in scientific fields. They ideally maintain continuous contact with reality. The raw data needs to be processed considerably to eliminate noise, callibration errors, intermittent sensor failure, etc. and then segmented and categorized. Various inferences can be performed directly on perceptual models, such as Kalman filters which predict what the world is going to look like at the next instant of time. Reactive AI and the availability of much better sensors and actuators has considerably advanced the state of the art in building perceptual models in the past decade.

- *Analog models* represent a (possibly hypothetical) situation in quantitative terms. Thanks to incredible advances in computer graphics and hardware may approach realistic conditions. Analog models can do inference through quantitative simulation, for example, using a physics simulator as used in computer games to simulate the execution of a recipe or the effects of gravity on objects and fluids.(Beetz et al. 2012) Sensors can be embedded in the simulation in order to answer questions about the course of an action.(Decuyper, Keymeulen, and L. Steels 1995)

- *Symbolic models* represent the situation in qualitative terms. Symbolic models are

less precise than analog or perceptual models and still leave a lot of the exact parameters for concrete action open. They already support inference to expand the model with defaults, contextual and background knowledge and using qualitative simulation. (Forbus 1988)

- *Linguistic models* are formulated in natural language or images and are primarily for communication among people. An example is the recipe text as found in a cook book. Linguistic descriptions are typically vague and highly underspecified and require a lot of knowledge, including knowledge of the language, to decode. For example, the cookbook may say 'add a handful of herbs' leaving it open which herbs to add or how much herbs constitute a handful. (Barnes 2004)

One of the main challenges in building understanding systems is to maintain intimate dependencies between these different models, namely to couple data from the perceptual models to the analog model so that the analog model becomes more realistic, to couple categorisations from perceptual or analog models to the symbolic models so that these models become grounded, and to couple the symbolic model to the language descriptions. Conversely, the language descriptions inform the symbolic model which constrains the analog model and provides expectations or completions of the perceptual model. Digital twins are AI systems that establish these dependencies during the life-time of an object. They are considered the basis for future smart manufacturing, system maintenance, retail, medical care, and many other fields.

Summarizing:

> **Digital twins are multi-layered models that construct and maintain a rich model of some aspect of reality at multiple layers: perceptual, analog, symbolic, linguistic). The utility of these models is based on how close they can represent and track reality.**

1.8 What are narrative-based models?

Traditionally, human-oriented disciplines (psychology, anthropology, economics, linguistics, semiotics, sociology, medicine, social neuroscience to name just these) characterize models as *narratives*. A narrative is a way to structure experiences. (J. Bruner 1991), (Vilarroya 2019) It identifies the relevant events connected principally by temporal and causal links, further enhanced with spatial, logical, analogical, hierarchical and other relations. A narrative identifies the actors and entities and the roles they play in events and the relevant properties of the events. It includes insights into the broader context and the motivations, deliberations and intentions of the actors. The more a domain is related to human issues, the more narratives also include a moral viewpoint on the events and an ideological framing.(Goffman 1974) So a narrative combines a set of seemingly disconnected facts and episodes into a coherent structure in which the relationships between

events and facts, the relevance of the events, as well as their moral and ethical framings are made explicit.

Narrative intelligence is the ability to come up with narratives, either based on observations of reality or on semiotic representations of narratives (texts, novels, drawings, movies) created by others. Narrative intelligence exploits general facts in *semantic memory* to fill in details, sensory-motor data stored in *episodic-procedural memory* to ground narratives into reality, *mental simulation* to imagine how situations in the world will unfold visually, and memory of past narratives, often called *narrative memory* - or *autobiographic or personal memory* if the narratives are about making sense of your personal life. In human intelligence these different types of memory are personal, based on your history of interactions with the world and others. Today knowledge graphs and other semantic resources are assumed to be universal, but we must expect that future meaningful AI systems have their own 'personal' dynamic memories which they have acquired through their own interactions with the world and other agents. They will undoubtly have their own opinions or ways of framing reality that may differ from that of other AI systems or from humans.(L. Steels 2020)

Narratologists make a distinction between three levels of narratives:(M. Bal and Boheemen 1997)
(i) There is the set of facts that the narrative is about. This is called the *fabula*. These facts may be objective data directly obtained by sensing the environment or qualitative descriptions without being colored by a particular viewpoint.
(ii) There is the *plot* which is the viewpoint, organisation and framing imposed on the fabula, partly in order to make a convincing narration.
(iii) There is the *narration* of the narrative, in the form of a text or another medium, for example as a documentary film, a theatre piece, a painting, an opera.[5]
Narrations use *signs*, which brings us on the terrain of semiotics. A sign is an association between a *signifier* (its form appearance) and a *signification* (its meaning) as governed by a *code*. The signifiers are constructed from material components (sounds, lines, colors, gestures, marks). Furthermore narrations abide by larger scale narrative structures classified and studied in the field of narratology.

Elements at the three levels of a narrative are intricately linked with each other. For example, a real world person in the fabula becomes a character in the plot and is presented by particular signifiers in the narration, such as a proper name, a specific dress, perhaps a melodic theme in an opera. Creating these linkages between levels is an important part of the understanding process and narrations have to contain enough cues to make these links detectable by readers or viewers.

The *narration* of a narrative by a narrator consists in gathering facts from observations or from collected data, selecting key facts, and organising them in terms of a plot, including the introduction of a viewpoint and framing of the facts, and then translating the plot

[5]Somewhat confusingly, a narration is also often called a narrative (cf. narrativo in Spanish), whereas here 'narrative' refers to the facts, the plot *and* its narration as a story.

into a text or other form of semiotic representation by choosing signs that introduce the elements of the plot. Conversely, an interpreter has to recognize and decode the signs, connect their various significations into a coherent representation of the plot, reconstruct from the plot the narrative and the underlying facts, and ground these facts into observed data. The interpreter also has to fit the reconstructed narrative into his or her past experience stored in a personal dynamic memory. Each of these processes is in itself extraordinarily complex. Understanding how they work and operationalizing them are the core technical challenge for the advancement of meaningful AI.

Clearly narratives have different purposes and characteristics and each narrative is a point in a continuum along many different dimensions. One important dimension is the *veracity dimension*, where on one side we find non-fictional narratives considered (or claimed) to be true in the sense of conform to reality. At the other end of the veracity dimension we find *fictional narratives*. They have ingredients based on reality but they do not pretend to be verifiably conform to reality. They may include fictional characters and events, exaggerations, unproven causal relations, rearrangements of the temporal order of events, etc. These changes are intended as the basis for narrations that are more compelling and hence more effective in convincing others of a particular viewpoint and ensuring that the narrative spreads faster in the population. Fictional or semi-fictional narratives arise spontaneously if not enough facts are known but people still try to make sense of what happens to have some degree of prediction and control over their world.

The human propensity for creating narratives and adopting, modifying and communicating them as narrations is particularly well illustrated with the discourse on the Covid pandemic and vaccination campaigns. Because there is a general lack of understanding and an ongoing change in the nature of the covid virus, we see that scientific theories need to adapt constantly to new insights, observations, the behavior of variants and the behavior of populations (keeping distance, masks, etc.). But we also see a wealth of spontaneous semi-fictional narratives, some of them taking the form of conspiracy theories or fake news stories, narrated through memes on social media, that are actually harmful for those believing them and hamper gaining collective control over the pandemic.

The various disciplines that use the concept of a narrative provide important insights into the nature of narratives and the functioning of narrative intelligence that are very useful to advance the state of the art in emulating understanding in artificial systems. These disciplines have also produced many case studies and challenges for concrete AI experiments which suggest new application areas for AI. Let me just give two examples.

Economics and sociology use the term 'narrative' in two ways, either for the narratives that people develop about economic and social phenomena, like inequality, or for the scientific models of these processes, in which case we speak about socio-economic theories. Scientific socio-economic theories strive for high veracity and typically treat humans as rational economic agents. 'Folk' economic narratives are often low on the veracity and rationality scale, but they can nevertheless have an important impact on socio-economic behavior. Identifying folk socio-economic narratives and studying their impact on the

real economy is the primary topic in the recently emerging field of narrative economics (Schiller 2019) which includes studies of the strategic use of narratives in order to advance economical agendas as is common in advertising or negotiation. There is also the emerging field of narrative sociology(Irvine, Pierce, and (ed) 2019), which focuses on how narratives shape social movements and social identity. It has similar concerns and approaches as narrative economics but focuses on social behaviors. Narrative sociology has strong interactions with social psychology and social neuroscience where the study of narratives and particularly narrative pathologies (leading to the construction and belief in conspiracy theories or radicalization) plays a growing role.(Willems, Nastase, and Milivojevic 2020)

In *medicine* there is a similar dichotomy between non-expert narratives and scientific theories. Patients spontaneously construct narratives to come to grips with the symptoms and evolution of their disease. Although many doctors do not feel the need to encourage or influence these narratives, there is a field called narrative medicine which sees the construction of such narratives (in a co-construction between doctor and patient) as a path to healing or maintaining health, complementary to natural science-based treatments.(Charon et al. 2017),(Sools, Tromp, and Mooren 2015) Narrative medicine encourages close (attentive) reading of texts, authentic discourse and reflective writing with an emphasis on personal reflections on facts.

Knowledge-based medical AI is concerned with the natural science-based approach to the study of diseases and treatments. In that field medical narratives are making headway as well in order to support search and hypothesis generation.(H. Kroll, D. Nagel, and Tilo-Balke 2020) The notion of narrative is more narrow here than in the social sciences and humanities and the term theory is often considered more appropriate. The narratives now focus almost exclusively on temporal and causal relations and recurring explanation patterns. Veracity and rational argumentation are primary and a moral stance and rhetoric to express this stance does not play a role.

Summarizing:

> **Narratives play a central role in understanding because they are the frameworks that provide the structures underlying the kind of rich models humans make. These frameworks are imposed on experience in order to bring out the relations between events and entities playing a role in them, formulate a moral stance, justify our actions and decisions, remember new experiences in terms of past ones, put experiences into a broader context, and communicate experiences and our views on them to others.**

1.9 How are narratives studied in other disciplines?

In order to further clarify the notions of meaning and understanding and the role of narratives, the MUHAI project decided to explore how other research disciplines traditionally concerned with meaning and understanding in domains of human interest have

approached the subject. These investigations should also give us ideas for which applications computational forms of narrative intelligence could make a contribution and how these opportunities might be realized.

The disciplines we consulted in some detail in the MUHAI project include social neuroscience and social psychology, economics and sociology, medicine, semiotics (especially art interpretation) and linguistics (especially pragmatics). The choice for these disciplines is partly based on the competences available with the MUHAI partners and partly to cover different points in the multi-dimensional space of narratives.

We already discussed the *veracity dimension*, referring to how true a narrative is targeted to be with respect to reality, in other words whether the narrative is about reliably representing and interpreting what has happened or about giving meaning to what happened.(Vilarroya 2019) The second dimension is about the role of *rationality*. On one end of the spectrum we find narratives that are based (or at least attempt to be based) on indisputable facts, rigid logic, and a view of humans as rational agents that maximally optimize their objectives. On the other end we find narratives that emphasize values and meanings and strive for a maximal fit with (personal) prior experiences. Next there is a dimension of *rhetoricity*. Scientific writing strives for clarity and coherence, whereas narrations developed primarily for convincing others may maintain ambiguity and even inconsistency which in some cases turns out to be more effective to persuade others.(Polletta and Ho 2021)

The different dimensions of narratives in social, medical and humanities research fields are summarized in the following table:

Dimension	Definition	High	Low
Veracity	Relation to reality	Truthful	Suggestive
Rationality	Argumentation basis	Logic and utility maximization	Compatibility with values and fit with prior experience
Sobriety	Persuasion style	Close to facts and logic	Amplified and selective expression of facts

Research disciplines can also be characterized in terms of the *topic* of the narratives which they are concerned with and the *facets* of a narrative they focus on. The topic ranges from the personal and the social to the natural world. The natural includes both the physical world, scientifically studied by physics and chemistry, and the living world studied by biology and medicine.

Here is now a brief discussion of the different research fields that have been considered by the MUHAI consortium, characterizing them along these dimensions. Each field is worked out in much greater detail in the next chapters of this report.

Economic narratives are analysed in Chapter 2 by Carlo Santagiustina. He points out how economic narratives are increasingly being studied as central to understand socio-economic behaviors of citizens by economists - with the recent Nobel prize for 'narrative economist' Robert Shiller as a sign how their importance is viewed in the field. The narratives most relevant to economic behavior have to do with inequality, technological innovations and products, or governance. They are spontaneously constructed by citizens in order to make sense of events in the world. These narratives are often constructed or manipulated by economic agents which have a high stake in how citizens make consumption or voting decisions. Hence they have a strategic purpose, which is most obviously the case for advertisement but also in advocacy of policies by politicians, anouncements by companies or governments or social media messages. Very clearly many of these narratives do not call upon rational argumentations and neither do they stay close to objective facts. What counts is their persuasive power that determines how fast they propagate in a population.

Historic narratives are analysed in Chapter 3 by Remi van Trijp and Inès Blin. Historical sciences, but also journalists or citizens trying to make sense of the past, face considerable epistemological difficulties. They have to make do with sparse, fragmentary, ambiguous, and uncertain input information. The output of historical sciences is also often in terms of narrations that are unavoidably biased towards the point of view of the historian (even if s/he is unaware of it) and amplify or downplay certain events in order to help the reader make sense of events. Van Trijp and Blin also emphasize a continuum of historical narratives from the particular, which sees an historical event as unique and to be described in full detail as such, to the instantiated (or embedded), which sees an historical event as a re-enactment of a recurrent pattern. They focus in their study on the French revolution which can be seen as a single particular event or as a manifestation of the recurrent pattern of all revolutions. In a similar way the invasion by Russia under the instigation of Putin against Ukraine today is seen by many as a manifestation of a recurrent pattern of aggression and war, similar to the invasion of Nazi Germany under the instigation of Hitler against Ukraine (and many other countries) in the nineteen-thirties.

Clinical narratives are analysed in Chapter 4 by Lise Stork, Ilaria Tiddi and Annette ten Teije. They are a way to understand causally the outcome of treatment effects through clinical trials. These narratives are built up collectively by the community of biomedical researchers and lately there have been important efforts to formalize these narratives. Clinical narratives are therefore a very good domain to study the formation and sharing of narratives with the support of AI tools. Stork et al. survey the tools currently being used and focus on important functions of deliberative intelligence needed for human-centric AI and made possible by narratives, namely hypothesis generation, explainability and perspective detection.

Social narratives are discussed in Chapter 5 by Oscar Vilarroya. Vilarroya comes from the field of social neuroscience and social psychology and in that field narratives have been shown to take a central role in human decision making, including for pathological cases such as radicalisation towards extremism. He compares the narrative paradigm that

dominates human communication and cognition with the rational paradigm that dominates scientific model making and makes the case for understanding narratives as having an adaptive function, namely to allow individuals to make sense of their surroundings and happenings and not so much to represent our experiences in a veradical way. Vilarroya then goes deeper into the cognitive and social processes that play a role in radicalisation, pointing to the existence of 'sacred values' and of various factors that influence in how far narrations impact the narrative making processes of individuals. These insights, backed also by neurophysiological data, are of great importance for the investigation of social and economic narratives and provide perspectives for dealing with the rampant negative effects of social media.

The remaining two papers focus more on the third pillar of a narrative, namely narrations. A narration is the way a narrator communicates a narrative by selecting specific aspects, not only the factual ones, and choosing a configuration of signs in a particular medium (for example language) that the narrator considers to be the most effective vehicles to get the key points of the narrative across.

Narrative art interpretation is discussed in Chapter 6 by Luc Steels. The field of semiotics has a long history of studying signs, how they are grouped as codes, how they originate and propagate and how they express narratives. The interpretation and production of art works, more precisely paintings, is one subfield of semiotics. In the case of art, the topics are mostly restricted to personal and social issues and high veracity and rationality are almost irrelevant. Instead effective use of rhetorical (expressive) devices is a major objective.

It is obvious that AI can learn a lot from semiotics about the kinds of narratives that human populations find important and how they are narrated, even though semiotics is seldom concerned with explicating the mental processes that allow humans to participate in semiotic exchanges. Significant recent advances in computational image processing and also the availability of a fast growing set of cultural facts in encyclopedic repositories is however leading to a new generation of tools for art interpretation(Klic 2021) making the integration of meaning and understanding in art a more realistic endeavour.

Narrative Pragmatics is discussed in Chapter 7 by Anna Morbiato. Language is the most complex sign system that humans have developed It is also the most studied system and a wide range of tools are available in AI to process language in the service of narratives although there are still significant gaps. The existing tools are both coarse-grained, i.e. based on statistical machine learning, usually grouped under the header of NLP, and fine-grained, i.e. based on linguistic theorizing in which case they are grouped under the label computational linguistics.

The area in linguistics that is most relevant for the purposes of dealing with narrations (in comprehension and production) is pragmatics. Pragmatics studies the linguistic markers available to steer the attention and support the narrative building processes of readers. Unfortunately there has so far been less work on operationalizing the insights from pragmatics than on the handling of morphology and syntax. Moreover the forays made

into semantics, for applications like topic modeling or sentiment analysis) are very coarse grained, ignoring most of the time information embedded in the language code.

We conclude:

> **Engagement with other human-centered research fields helps the development of meaningful AI by showing the kinds of narratives humans develop, and how these narratives are constructed, remembered and exchanged through narrations.**

The following sections of this report contain the more detailed discussions for each research discipline, followed by a concluding section identifying bottlenecks and opportunities for advancing the technology of human-centric AI.

Acknowledgement The author thanks Oscar Vilarroya (IMIM and UAB Barcelona), Lise Stork (Vrije Universiteit Amsterdam) and Katrien Beuls (Université de Namur) for valuable comments on this paper. The author is also indebted to many discussions with other members of the MUHAI consortium, particularly during meetings with Frank van Harmelen and his team at the Vrije Universiteit Amsterdam.

Chapter 2

From Narrative Economics to Economists' Narratives

Carlo Santagiustina

Abstract

In the last decade, there has been an increasing interest among economists, policy advisors and social scientists for narratives, and for their role in relation to consumers' and citizens' decision-making processes, public agenda-setting, policy making as well as its explanation and justification to the general public. In particular, a growing body of literature shows that narratives related to socio-economic issues of collective interest, like inequality, play a concrete -if not crucial- role in the evolution of observed socio-economic phenomena and related policies. In this work, we review literature from economics and neighboring fields to identify key differences among alternative approaches to the study of narratives. Second, we highlight the strategic role of narratives and their ubiquity in relation to the different phases of agents' decisional processes. Finally, we discuss the relation between narratives by academic researchers and narratives that spread among the general public or on the media. This, to appraise the endogeneity of social and economic research activity in socio-economic narratives dynamics, and to better understand their role in governance systems. This work suggests that the power of narrative extends well beyond their observed virality in the news and in social media, and that narratives' impact is a foundational aspect of humans' way of making sense of their surrounding reality, also in the economic domain.

Keywords

economics, narrative economics, economists' narratives, economic discourse

2.1 Introduction

Seeking the roots of narrative economics. It is no accident that the expression *narrative economics* was used more than a century ago in Palgrave's *Dictionary of Political Economy*, to refer to a method for constructing an economic explanation of historical events. Interestingly, in the first Volume of this dictionary, under the heading of `EXPERIENCE` (Palgrave 1894, p. 790-791), it is stated that:

> "*An economist strictly deductive in method could never get beyond his first premises. The contrary seems possible because the economist who apparently deduces everything from first principles in reality weaves into his argument statements of fact and wide generalisations which have become so familiar that he and his readers forget how they were first acquired. Nor has there ever yet existed an economist who merely observed and recorded. [..] The writing of history involves processes of selection, comparison, and inference, in which the historian's mind is active. No two persons perform these processes in quite the same way [...] It is not merely that historians often infuse their work with their own political or religious sentiments, with the prejudices of their own age or their own class. It is rather that the historian cannot construct a narrative out of facts without interpreting those facts. But he cannot interpret the facts without using his mind, without adding to, or rather, without transforming, those facts.*
> *The object of all science, including political economy, is not merely to amass facts but also to explain them.*"

As Palgrave's words suggest, economic narratives are representational constructs that humans incessantly construct, employ and transform for explaining facts, for reasoning about economic events of collective interest, and about their patterns across space and time. As we will try to illustrate through this work, narrative's generalizations mentioned by Palgrave are, among others, necessary for rapidly categorizing and making sense of what we experience on a daily basis as individuals and groups that operate through complex market and socio-economic systems, and by so doing reduce uncertainty. As such narratives play a significant role in aggregating, and (possibly) learning from, common experiences and recurrent and seemingly related sequences of economic events. Through their rhetorical constructions (Phelan 1996; Kirkwood 1992; Rodden 2008), narratives select, combine, (re)frame, and historicize mixtures of events, actions, choices and preferences that are believed to depend on one another. Despite economic narratives tend to be (more or less) homomorphic to the evidence and relations they describe, when recalled, they can distort and blur our experiencing of the economic world (Keusch, Bollen, and Hassink 2012), while projecting sensory experiences and data into the realm of cognition, via natural and formal languages. Narratives are intermediate cognitive (and meta-cognitive) goods (Scholz 2018), which can be analyzed, used, transformed, transferred, and conserved for a multiplicity of practical or epistemic purposes. For example, stock market narratives in specialized news can be used by investors to rapidly interpret the causes of recent changes in market trends and react accordingly, or to predict the likelihood of the intervention

of the Federal Trade Commission in a merger between firms which could damage consumers, or to persuade EU citizens and investors that a the a non-standard monetary policy employed by the European Central Bank will succeed in limiting inflation to its targeted value, given a claimed set of dependencies linking the newly employed policy levers to inflation. Narratives can be more or less economically relevant and impactful, depending on their diffusion, content and persuasive power, which is often derived from their rhetorical strategy and believability, the authority of their declaimers and supporters, their fitness to a scope, the salience of their message or implicit recommendation with respect to the situation in which, and audience with whom, they are employed for informing or conditioning decision-making processes.

The process of constructing a narrative out of facts. Narrativization is the mechanism through which human beings chain their individual and collective experiences together through language (White 1981). Not only from a temporal perspective but also from a causal point of view. This activity can be done at the individual level, for example, to make-sense of a higher than expected electricity bill, or at the group level, for example, to convene on a narrative for explaining to European consumers today's EU reliance on natural gas imported from Russia. Clearly, languages through which economic narratives are built encompass natural languages, but also embrace applied mathematics (Mary S Morgan 2001), statistics (Dumez and Jeunemaitre 2006) and all other symbolical representations of the real-world and of its phenomena, including (symbolic) behavior (Brown 1986).

Economic narratives in action. As shown in Figure 2.1, which displays a Twitter post about the relation between the facial expression of the photographed trader with upcoming NYSE trends, even a simple tweet can combine different mediums and visual languages for creating a unifying (re)construction of specific situations or events, and of their relation, also called a narrative.

As observed by (Zaloom 2003), "*the body language of the trader, who may be steadily and confidently holding his hands forward in engagement with the market or yelling his bids, spittle flying and eyes wide, in desperation to get out of a trade, are crucial inflections that traders draw on to form market judgments*". Regardless of its unconditional veracity, if this narrative is sufficiently diffused in the traders community and "*inflections*" (like the one shown in Figure 2.1) are considered crucial for forming market judgments by market operators, then, as in a self-fulfilling prophecy (Merton 1948), the relation described by the narrative may well become a (conditionally true) mechanism driving the market. As for the Keynesian beauty contest, in the economic world, it is the belief in, and coordination through, a narrative that often makes its claims become true (Tuckett and Tuckett 2011).

Figure 2.1: Examples of stock-market related Twitter post containing a narrative, claiming that forthcoming market trends of the NYSE can be "known" by looking to one of the most photographed traders (Peter Tuchman) depicted in the image attached to the tweet. Tweet link: `https://twitter.com/Albannay/status/978380005087633409`

Purposes and contexts of narrativization. Even though the reasons for which narratives are created, employed and shared in real-life situations are innumerable, some of their usages are more relevant than others to economists and other social scientists. In this respect, there has been a recent focus on how narratives can (i) prime and nudge individuals and groups (Heath, Lee, and Lemon 2019), for example, in relation to their goals (Laham and Kashima 2013); (ii) frame decision- and sense-making processes (Steinhardt and Shapiro 2015; Hullman and Diakopoulos 2011); (iii) influence memory recall (Thorndyke 1977) and information salience (Dolan et al. 2012); and (iv) coordinate group behaviors, identities, and declared preferences (Lebano and Jamieson 2020).

In liberal and democratic societies, where the justification and explanation of individual and collective choices is a relevant aspect of daily life and deliberations (Boswell 2013), narrativity has become the (default) means through which people's representations, expectations and choices are coordinated via horizontal communication.

Narratives are especially important in decentralised or non-hierarchical systems, like markets and social media, because coordination and alignment cannot be therein imposed. As

a result, narrative constructions have become fundamental ingredients in those processes whose success depends on voluntary coordination and compliance with recommendations. Like vaccination choices (Abeysinghe 2015; Poltorak et al. 2005) or the adoption of socially and environmentally sustainable behaviors and investment choices. Moreover, since narratives can affect expectations (Svetlova 2021), and expectations can stabilize or destabilize economies and drive major events (G. W. Evans and Honkapohja 2012), one would expect them to be subject to close scrutiny by economists.

Why have narratives been stigmatized in economics? Regardless of their relevance, until recently, works focusing on economic narratives have been mostly neglected by economists and top ranked economic journals, with few exceptions like Ahern and Sosyura 2014; Ashenfelter 2012; C. D. Romer and D. H. Romer 2004; C. D. Romer and D. H. Romer 2010. I hypothesize that this is mainly due to two reasons, illustrated here below.

First, narratives have been often stigmatized and considered a taboo in the economic field because they are seen as being related to subjectivity, interpretative arbitrariness, measurement difficulties, and the impossibility to derive (universal) economic laws from evidence. Especially in economic departments culturally akin to the so-called *Laissez Faire* philosophy, researchers have been taking part to a coordinated effort to generalize, extend and test neoclassical economic models and their assumptions. Possibly seeking the perceived scientific authority (Bijker, R. Bal, and Hendriks 2009) granted to natural sciences (Copeland 1931), like physics (Pikler 1954; Pikler 1955; Walker 1991; Mirowski 1991) and engineering (Ekelund et al. 1999), different generations of *Chicagoan* researchers (Miller Jr 1962) contributed to the building of the edifice of contemporary economics, mostly through accrual and non-disruptive contributions to the neoclassical paradigm. Such a theoretical edifice, by being mathematically sound, would hypothetically allow the economic discipline to ensure the benefits granted by the so-called hard science approach (Godechot 2011; Keen 2001), including status and research funding (Fourcade 2009). As a result, during this quantitative turn where every scientific dispute had to be resolved by formal demonstrations or statistical tests, the idea of studying seemingly soft issues (G. A. Akerlof 2020), such as narratives, was probably considered an inconvenient or too risky move to most economists, which are by training accustomed to assess career incentives and funding prospects conditional on the research paths they undertake (Harris et al. 1959).

Second, acknowledging the role and effects of narratives in economic processes and outcomes would have been equivalent to admit that (also) economic research, its policy implications and related narratives can affect (for the better or the worst) the systems that they describe; with all the resulting socio-political and economic implications and responsibilities, which academic researchers most likely prefer not to confront with.

As a result, even though narratives have always been known (unknowns) in economics, before the pioneering work by Deirdre McCloskey (McCloskey 1983), narratives were

broadly considered marginal issues, transient noises, incommensurable or unobservable variables, or negligible 'Blah Blah Blah'; whose meaning could not be deciphered given the lack of methods for identifying, extracting, interpreting, aggregating, and modeling them. By neglecting the role of narratives in economics, economists were implicitly assuming that narratives (of any type) couldn't significantly affect the state of economic and financial systems they were studying, nor influence their dynamics. As we will see, this extremely strong assumption has been proven to be incorrect through many economic applications.

The digital revolution: a precondition for an empirical narrative turn? In the last two decades, the rapid growth of the World Wide Web and the progressive digitization of our socio-economic life (DiMaggio et al. 2001), together with the theoretical and methodological advances in knowledge representation (Van Harmelen, Lifschitz, and Porter 2008), cognitive linguistics and AI (Lippi and Torroni 2016; V. Evans 2009; Duchan, Bruder, and L. E. Hewitt 2012; Luc Steels 2011), have transformed discourse and narratives into latent evidence, which, with the appropriate techniques, can be captured and transformed into standardized information structures, and used for modeling, forecasting (or simply describing) socio-economic systems and their information environments.

During this digital shift, there has been a growing attention, also in economics, to stories (R. Akerlof, Matouschek, and Rayo 2020), narratives (Robert J. Shiller 2017), public discourse (Ferrara et al. 2021), and their effects in real world economic and social systems. Through this growing body of research, it has been found that narratives are important in relation to policy making (Cherif, Engher, and Hasanov 2020), justification (Gaur and Kant 2021), legitimization (E. Hewitt 2020), and compliance (Mintrom et al. 2021). This attention has also prompted the emergence of *narrative economics* (Robert J Shiller 2020) as a new paradigm for doing socioeconomic research, by focusing on people's communicated views about the functioning of economic systems, the determinants of economic outcomes, and the causes and consequences of economic events. This research has also raised questions regarding the methods, evidence-collection procedures and modeling paradigms that should guide economists and other social scientists in integrating narratives and discourses into their research fields (Meng 2021).

Structure of the work. The sections that follow provide an overview of the emerging field of narrative economics, and of economic narratives' life-cycle and ecosystem. Highlighting how they are crafted and used to influence sense- and decision-making processes, at the individual and collective level. In section 2.2, I review how different research streams have defined and studied narratives in relation to economics. In section 2.3, I further explore the role of scholars in the generation and diffusion of economic narratives. Finally, in section 2.4 I summarize the most relevant implications of narrative research in economics and human-centric studies.

2.2 Economic narratives: tying societal discourse and economic decisions

"If you listen to the noise of the market, you won't buy anything." (African proverb)

Characterizing economic narratives. Even though the word *narrative* is often (mis)used as a synonym of *story*, *anecdote*, or *discourse* in economics, from an epistemic perspective narratives are very different from the latter, mainly because, as pointed out by (Boswell 2013), they operate at an intermediate level between personal anecdotes and public discourses, constraining agency by limiting the possible ways of viewing an economic issue or knowing an economic event.

In particular, narratives are systems of purposefully selected and framed (anecdotal or factual) evidence, which, by being causally and sequentially related in a plot, contain emergent meanings and abstractable teachings, which suggest a specific stance or (re)action in relation to the described economic issue or event. The *raison d'être* of economic narratives is not merely to describe something related to the world of economic affairs, but also to suggest something to do, or to think, in relation to what is described. Strategically designed economic narratives can serve as understanding recommendations (or knowledge primers). Narratives can be instrumental to the diffusion of non-exhaustive representations of (economic) evidence, tailored to meet the "*doxastic preferences*" (Baltag and Smets 2013) of specific individuals or groups. For this reason, especially in economics, narratives should be always analyzed as means rather than ends, which can be employed, when considered convenient to do so, for justifying, motivating, explaining or encouraging specific ways of thinking, behaving, and expecting.

Processes of narrativization imply a (non-casual) selection of situations and events, experienced by oneself and/or others, to be linked in a plot, as Figure 2.2 shows, different narratives can be constructed from the same evidence set. The way in which situations and events are selected and linked together is often conditional on, and instrumental to the end purpose for which a narrator creates or spreads a specific narrative. For example, the plot in Figure 2.2 could be used to explain why a general consultation resulted from specific choices, situations and events through which a group of individuals had to do with each other, in a (seemingly) strategic way. However, the process of "*cherry picking*" evidence (Fox and Hoch 2005) for building a narrative, even if carried out strategically (Miskimmon, O'loughlin, and Roselle 2014; Van Noort 2017; Freedman 2015; Miskimmon, O'Loughlin, and Roselle 2015), is not necessarily rational (Elster 1979), as it may be affected by heuristics and biases (Tversky and Daniel Kahneman 1974; Daniel Kahneman et al. 1982) that are typical of judgment under uncertainty, like the anchoring, adjustment, representativeness, and availability heuristics. Many of the heuristics that have been found to influence decision-making can also influence the construction of narratives. This is important because if cognitive biases, possibly affecting in a negative way decision-makers, can be first identified by studying narratives, irrational decisions could

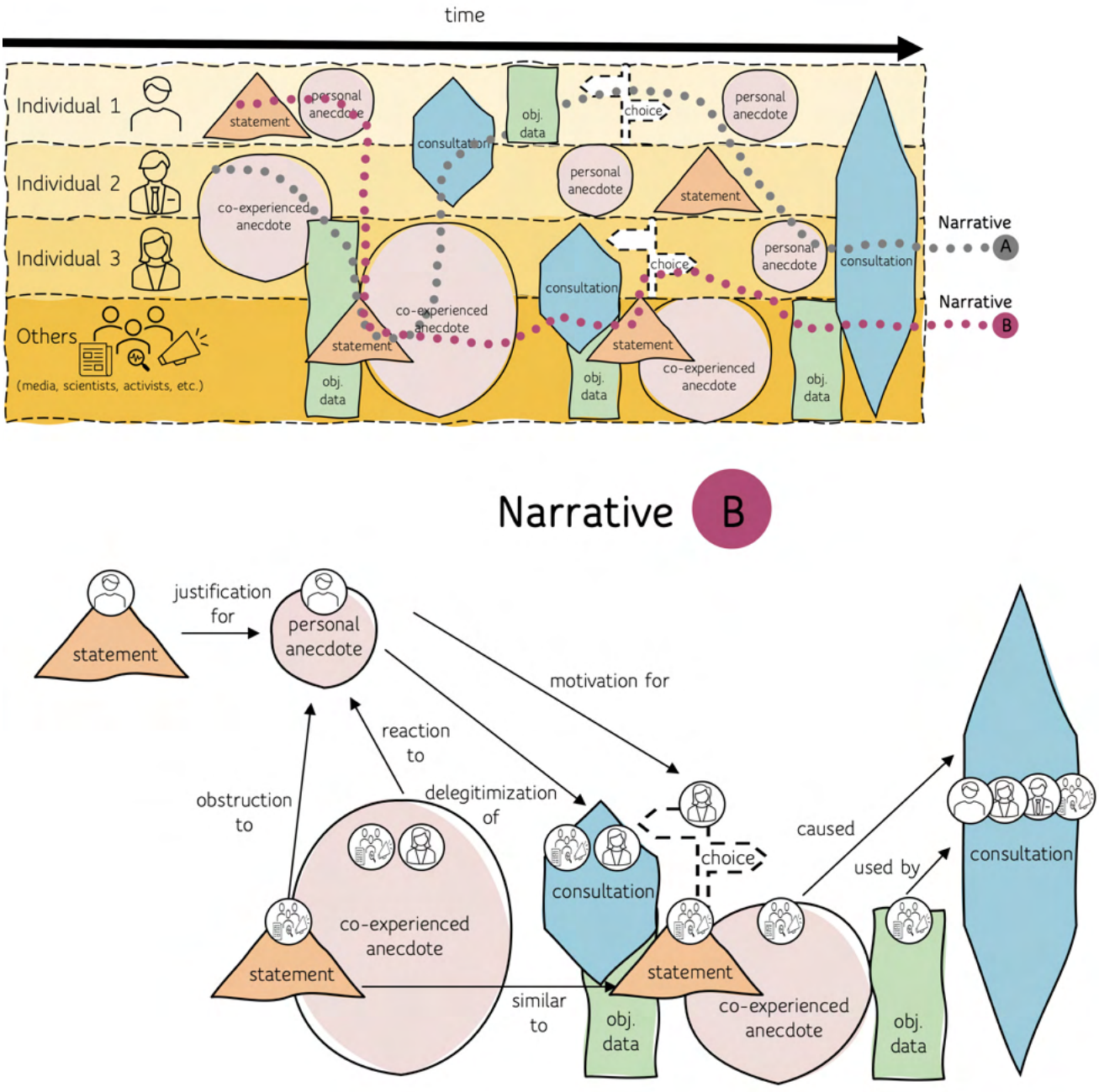


Figure 2.2: Narrativization processes and emergent narrative structure

be possibly anticipated or avoided. The relation between cognitive biases revealed through narratives and biases in decision making, opens the way to a promising research area, that of the forecasting of (individual or collective) behavioral biases through the analysis of narratives and their dynamics. For example, one could try to anticipate irrational col-

lective behavior by looking to the dynamics of narratives about collective risks that may be socially amplified through communication on social media (Iacopini and Carlo R.M.A. Santagiustina 2021).

Narratives as uncertainty avoidance tools. As explained by (Tuckett and Tuckett 2011), narratives can be seen as tools for reducing uncertainties by constraining people's freedom of representation:

"*The word narrative has two etymological roots – telling (narrare) and 'knowing in some particular way' (gnarus). The two are so intertwined they cannot be untangled. Bruner (J. S. Bruner 2003) uses this point to summarise the importance of narrative particularly for giving us ready and supple means for dealing with the uncertain outcomes of our plans and anticipations – citing Aristotle to note that the impetus to narrative is expectation gone awry – peripeteaia, or trouble. Among other possible functions, narratives provide a vocabulary of meaning to support and legitimate action and to deal with misfortune.*"

By (self-)limiting the possible ways of viewing the world, shared narratives that are sufficiently stable across time can give relief to the overwhelming process of contemplation of the complexity and variety of real world experiences and perceptions, which, by inspiring a multiplicity of inconsistent feelings and representations, may otherwise give rise to uncertainties (Carlo R.M.A. Santagiustina 2018), to which individuals are averse to.

To reduce uncertainty, people can use (individually and collectively) their imagination, culture and episodic memories to anchor their world-views and expectations to existing structural constraints, in a way that is instrumental to the fulfillment and justification of their preferences and objectives (see Figure 2.3). These representations and expectations are then put to the scrutiny of the community or/and co-deciders in the form of narratives, which jointly explain and justify a plan of action and its (claimed) end(s). By so doing, narrativized relations between past, current and future situations are shared in social networks and hence embedded into the socio-cultural context of the decision-makers. Under this perspective, via social, cultural and personal frames, narratives bind possible futures to contingent structural constraints.

Based on these socially-constructed representations people may feel more comfortable in taking decisions in a condition of objective uncertainty. These decisions may both consist in taking or postponing choices and actions, in accordance to the unfolding of the narrativized "*plan of action*". As a result, narratives allow people to act "*according to or despite the uncertainty they face*" (Vignoli et al. 2020).

Collective Economic Narratives. In order to solidify the foundations of narrative economics, a definition of Collective Economic Narratives (CEN) has recently been proposed in (Roos and Reccius 2021). According to this definition a CEN is "*a sense-making*

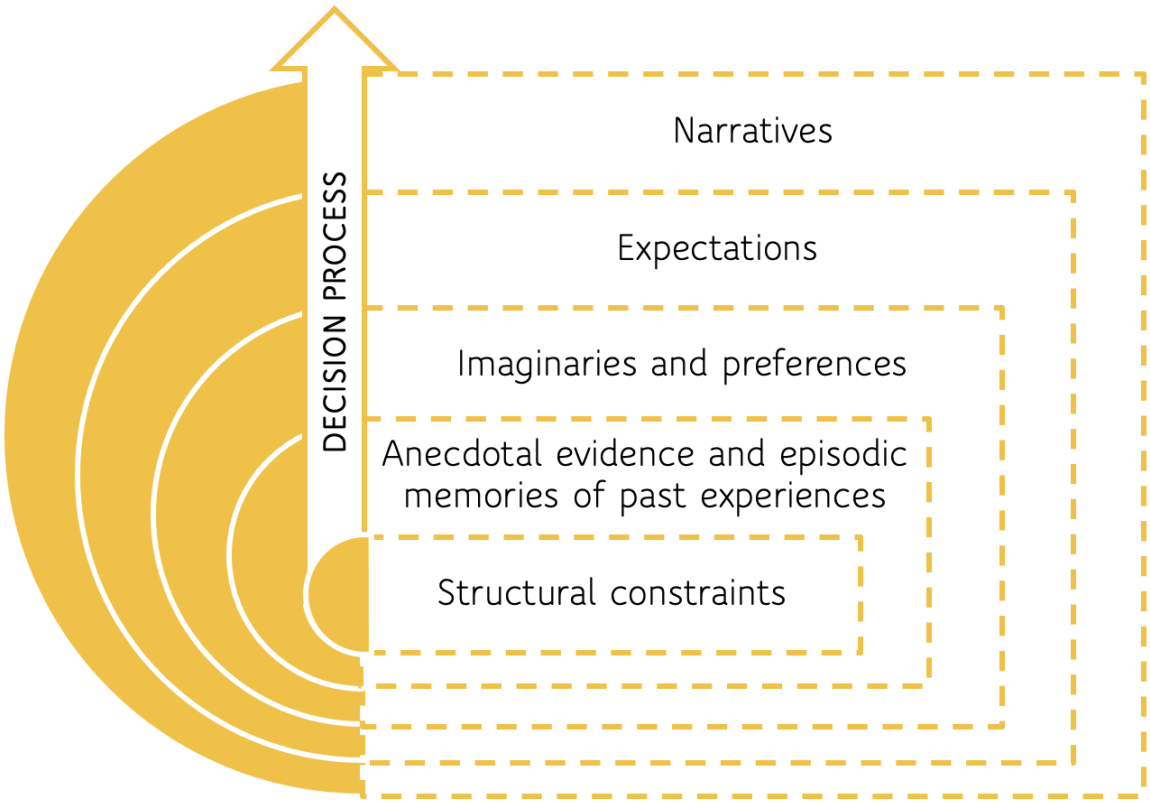


Figure 2.3: the Narrative Framework, image inspired from Fig.2 in (Vignoli et al. 2020)

story about some economically relevant topic that is shared by members of a group, emerges and proliferates in social interaction, and suggests actions".

Despite the arguable (and loose) usage of the word *story*, the above definition highlights some important functional and substantial aspects of narratives, as seen from an economists' perspective, which are summarized in the box here below.

Collective Economic Narratives (from Roos and Reccius 2021)

1. ARE SENSE-MAKING TOOLS: "*In a social context, sense-making can mean that a justification for behavior is given. [...] Sense-making requires that the story connects to the belief systems of the involved people. We use the term belief system in a broad sense here, including mental models and normative, evaluative, affective and motivational elements*".

2. LINK ECONOMIC PHENOMENA: "*Economists are not interested in language per se, but only with regard to its relation to their main objects of inquiry such as the economic activities of production and consumption*".

3. CONTAIN LATENT MEANINGS: "*A narrative may transport subtext in addition to what is said explicitly and directly. The subtext appeals to the underlying belief system [...] their meaning and the relation to other topics are just implied and must be completed by the listener*".

4. EMERGE THROUGH COMPLEX INTERACTIONS: "*Nobody who tells a story to others can know exactly which parts of the story make sense to them and which do not [...] In the end, a group narrative is left that nobody thought of in this way and that nobody could predict, hence it emerged*".

5. ARE SUBJECT TO EVOLUTIONARY PRESSURE: "*The narrative of a group might be challenged by members of other groups that maintain different belief systems. Different groups [...] may have an incentive to differentiate their narratives [...] The evolution of a group narrative depends in complex ways on the participants and the rules and practices of the inter-group discourse*".

6. ARE USED TO SUGGEST OR FACILITATE ECONOMIC ACTION: "*Suggesting what to do in an uncertain world is another interpretation of sensemaking [...] people often have a desire to act, even though the knowledge basis for rational decision-making is rather small [...] This idea has some similarity to Keynes' concept of animal spirits – defined as spontaneous urge to action rather than inaction*".

As identified by Roos and Reccius, economic narratives serve to form behavioral scripts and cognitive categories at the group or community level that facilitate coordination, as well as sense- and decision-making processes. These scripts and categories are not necessarily based on objective reality, and may be the product of experiences and impressions that are subject to negotiation and change within a group or at the community level.

Through a case-study on the role of narratives in the Soviet Union (G. A. Akerlof and Snower 2016), Akerlof and Snower come to similar conclusions. In their view, narratives play a key role in focusing attention and activating memories and motivations, especially at the collective level. According to the two economists, these shared meaning constructions, by providing simple mental models of causal and social relations, can establish, maintain or change power relationships, also producing a strong influence on economic and political decisions. In particular, since people make choices with regard to a domain of possibilities that lies within their field of attention, and narratives act as attention filters at the societal level, they constrain people's domain of perceived possibilities, informing, guiding and limiting action at the aggregate level.

The aforementioned approaches to economic narratives are closely related to complexity economics literature, in particular to an emerging stream of research which sees the economy "*not as a system in equilibrium but as one in motion, perpetually 'computing' itself – perpetually constructing itself anew*" (Arthur 2014).

Research on CEN, with a focus on socio-economic inequality. Since the late 1980s, a growing body of economic research has started investigating the effects of narratives on economic phenomena, and their role in the formation of preferences and justification of economic policy and experts' recommendations. Research fields that were affected by this narrative turn include development studies (Palvia, Baqir, and Nemati 2018; Lewis, Rodgers, and Woolcock 2014; Gatrell and Reid 2017; Titumir 2021; Chan 2014), economic history (Ferguson-Cradler 2021; Eichengreen 2012; Kitromilides 2013), decision theory (Jerome Bruner 1990; G. A. Akerlof and Snower 2016; G. A. Akerlof 2020; R. Akerlof, Matouschek, and Rayo 2020; Singer 2015), microeconomics (Preece and Kerrigan 2015; Safari and Thilenius 2013; Hume and Mills 2013; Morrell and Jayawardhena 2010), macroeconomics (Tilly, Ebner, and Livan 2021; Robert J. Shiller 2017; C. D. Romer and D. H. Romer 2004; C. D. Romer and D. H. Romer 2010; Rojas, Vegh, and Vuletin 2020), finance (Hall 2006; Nyman, Kapadia, and Tuckett 2021; Teeter and Sandberg 2017; Iacopini and Carlo R.M.A. Santagiustina 2021; Costola, Iacopini, and Carlo R.M.A. Santagiustina 2021a), environmental and sustainability economics (Van Der Leeuw 2020; Béné et al. 2019; Bauer et al. 2017; Saltelli et al. 2020; Maller, Horne, and Dalton 2012). Given the scope of this work, we will here focus on development studies related to socio-economic inequality. For a more general overview of the role of narratives in other economic fields and related empirical works, we refer the reader to (Roos and Reccius 2021; Robert J Shiller 2020; Ferguson-Cradler 2021).

In development studies, it has been found that narratives can play an important role in determining the perception of and tolerance to different forms of economic and social inequalities (Gaur and Kant 2021; Larsen 2016).

For example, Gaur and Kant (2021) have identified and described a set of channels through which economic narratives can affect the implementation and impact of economic policies, like wealth redistribution policies, and alter people's wealth distribution preferences, possibly leading to further concentration of wealth and power. According to the aforementioned authors, institutions, media, and technology play a key role in creating, validating, or disseminating narratives that justify (or condemn) specific forms of socio-economic inequality. Hence, their impact is not neutral, and they can indirectly exacerbate inequalities, creating what the authors call "narrative-driven inequality". The concentration of wealth can be too such an extent facilitated by collective narratives, that under some conditions, these can determine a perpetual concentration of wealth through self-reinforcing mechanisms (see Figure 2.4).

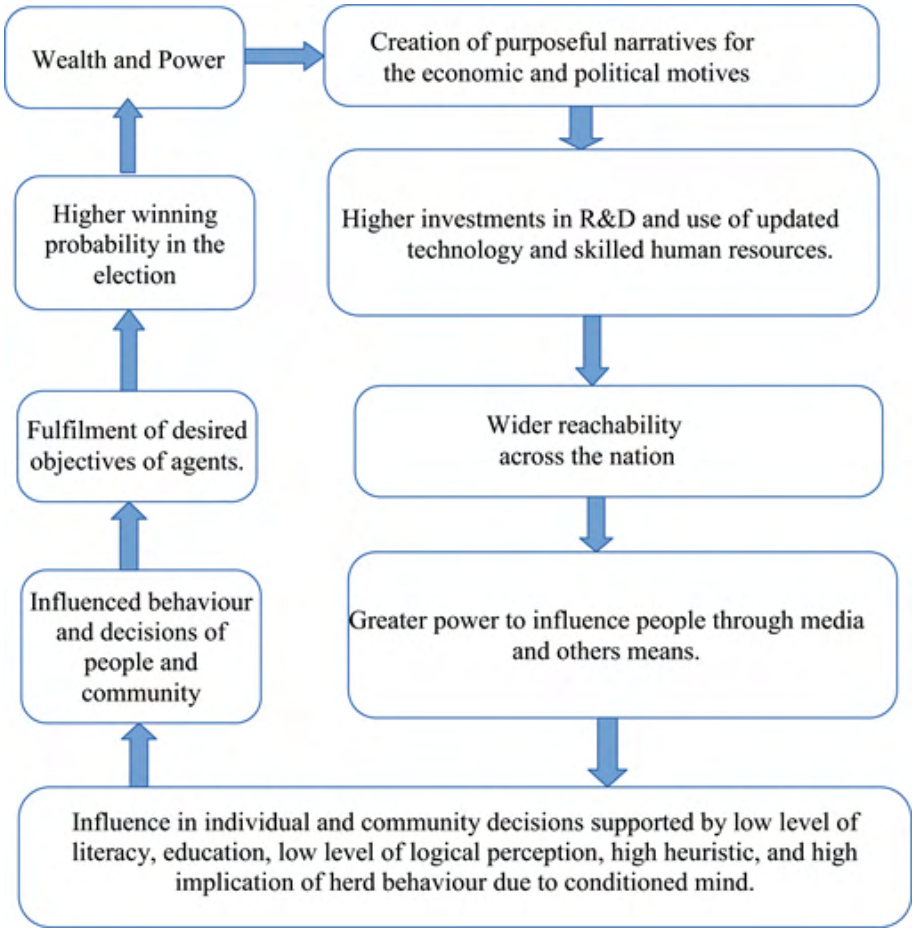


Figure 2.4: Linkages among wealth, power, narratives, media, and peoples' perception, image from (Gaur and Kant 2021)

2.3 On the role of economists' narratives: (Re)defining the boundaries of Narrative Economics

It has been claimed (Nakai 2020, p. 6), that economists' growing interest for narratives mirrors a progressive shift of economics to "*the more relational methodology of historical, social, and behavioral economics*", which would encourage further interaction and integration with neighboring fields in the social sciences, humanities, linguistics and AI.

Here follows a critical analysis to this view, and an attempt to explain why economics "armed" itself with a a new stream of research on economic narratives, called *narrative economics*. Rather than simply mirroring a rapprochement of economics to neighboring fields, the narrativization of economics and the emergence of *narrative economics* are above all the product of the growing influence of economic research on the political and economic system. This because, as we will see, academic economic research is endogenous to economic narratives and to the functioning of economic systems.

On the non-separability between economics and economic narratives. Economists' endogeneity in the systems they describe and model through their works (Herrmann-Pillath 2008; Andrikopoulos 2013) can be seen as case of omitted variables, possibly affecting the micro-dynamics of economic systems, the evolution of aggregate economic variables, as well as the design of economic institutions. For example, it has been observed that economists' recommendations and narratives in favor of austerity measures, before and during the European Sovereign Debt Crisis (2010-2012), have influenced its duration and outcomes (Crouch 2011). Not only economists can affect policy-making, like IMF's intervention timing and choices during the crisis, but also, they can play a key role in the formation of market operators' expectations, for example, by diffusing or acknowledging those narratives that claimed that strengthened austerity measures were necessary, though not sufficient, to avoid the bankruptcy of those EU countries that had accumulated a high level of public debt, such as Italy and Greece. This point is particularly important because, although economics aspires to be a positive science, its works and narratives often end up producing normative effects, pushing the economic-world to fit to economic theories, rather than the other way round, like a foot forced to mold its shape to that of a rigid shoe, and the consequences of these narratives are not always insignificant or painless. Some economists, like Sedláček Tomáš (Sedláček 2014), have gone so far as to claim that neoclassical economics and its axiomatic approach, is a meta narrative told in a formal language:

"*[T]here is at least as much wisdom to be learned from our own philosophers, myths, religions, and poets as from exact and strict mathematical models of economic behaviour. I argue that economics should seek, discover, and talk about its own values, although we have been taught that economics is a value-free science. I argue that none of this is true and that there is more religion, myth, and archetype in economics than there is mathematics.*"

These meta narratives operate at a more abstract narrative level, transforming economists' world-views into model specification choices, and economists' preferences over world-views into assumptions, variables and case-study choices, exploiting as justifications for the latter the seeking of "clarity","simplicity","coherence", "elegance", "intutiveness" and/or "parsimony".

The increasing discoursivisation of economic events, phenomenon that accelerated after the 2008 financial crisis and more recently during the COVID-19 pandemic, signals a slow but relentless change in economic thinking, towards a more pragmatic and narrative-centric understanding of economic affairs and of their dynamics. In particular, the (re)birth of narrative economics coincides with the revival of a debate among economists about which are the boundaries of economics, and urged many researchers to consider whether these boundaries are destined to disappear or, on the contrary, whether they should be walled to prevent trespassing, in both directions, with bordering disciplines and research fields.

In the last years, the "narrativisation" of economics has found its advocates also in economic departments, especially among behavioral and cognitivist economists, who crept into the fissures and unexplored areas of the unfinished building of economics, to experiment with new ways of understanding economic phenomena, by identifying (and eventually acknowledging) the role of narratives (Robert J Shiller 2020; Robert J. Shiller 2017). In addition, the digital revolution and the incessant emergence of huge amounts of textual data generated by (or in relation to) user behavior on online platforms, search engines and social media (Costola, Iacopini, and Carlo R.M.A. Santagiustina 2021a; Iacopini and Carlo R.M.A. Santagiustina 2021; Costola, Iacopini, and Carlo RMA Santagiustina 2021b; Carlo R.M.A. Santagiustina 2018), has opened a debate (Meng 2021) about which types of data should be considered by economists and other social scientists, when having to model or explain a phenomenon of interest, like macro risks perception, investor uncertainty or inequality aversion.

2.4 Conclusion

As illustrated through this work, in many cases, narratives have been found to be relevant factors capable of significantly affecting individuals' choices and events' outcomes, also in relation to market phenomena and economic issues. In specific situations, where economic outcomes and choices are poorly explainable or predictable if one considers only objective and non-narrative factors, like official economic statistics, discourse and narrative dynamics related to economic happenings often reveal to be valuable resources to understand and anticipate economic systems' dynamics, especially at the aggregate level.

Narratives are not simply lagging or biased representations of the real-world and its changes but also cognitive goods employed for sense- and decision-making purposes, as such they are a key determinant of our socio-economic reality and its dynamics. Popular economic narratives, even when not grounded on evidence or sound deductive reasoning, should be carefully considered by policy makers, social science researchers and economists, to understand to which degree observed economic events and revealed preferences are (or not) the by-product of the distribution of Collective Economic Narratives, rather than the outcome of an optimization process based on exogenous preferences and objective measures. This has relevant consequences both for economics and for economic research. Moreover, the existence of a dependence relation between (cognitive) biases revealed through narratives and biases in decision making, opens the way to innumerable extremely promising research areas for economics and neighboring disciplines, in particular, that of the modeling and forecasting of behavioral biases through the analysis of the dynamics of economic narratives, such as those related to social inequality and its causes. Finally, narratives will probably play an increasingly relevant role in economic phenomena, especially for what concerns immaterial and digital goods and their consumption, therefore their analysis is of utmost importance for the understanding and modeling of

online economic phenomena and related investment and consumption behaviors at the aggregate level.

Acknowledgements

I wish to thank the MUHAI family (https://muhai.org/people) for their precious feedback and inspiring discussions on narratives and related issues addressed in this work. I am extremely grateful to Luc Steels for giving me the opportunity to contribute to this foundational deliverable and for encouraging me to think critically and holistically about the role of narratives and rhetorics in economics and economic affairs; reminding me not to lose sight of a fundamental aspect of this ambitious project, its focus on the human beings, and on their use of language for (re)constructing and understanding the ecosystems in which they are embedded.

Chapter 3

Narratives in Historical Sciences
Remi van Trijp, Inès Blin

Abstract

Historical sciences such as geology, evolutionary biology or history try to offer causal explanations for non-recurrent phenomena (e.g. how the Grand Canyon was formed, how the human eye evolved, or what caused the Second World War), typically using incomplete and fragmentary evidence from the past. Even though these sciences make use of general frameworks such as the Theory of Evolution, they have to work out the specifics of each case and hence they cannot simply apply general laws and deductive reasoning. Instead, causal explanations are expressed in narrative form. While such explanations have long been considered to be "less scientific", there is now a growing awareness that narrative explanations go beyond mere description and also have the potential for empirical testing. This Chapter explores how narratives are used by historical scientists, and how human-centric AI systems may assist scientists in constructing more precise and testable narratives so they can achieve a deeper understanding of society. It presents a first prototype that takes as its case study the French Revolution (1789–1799).

3.1 Introduction

Narrative explanations play a major role in historical sciences such as geology, evolutionary biology, and history that have to provide explanations for non-recurrent phenomena based on evidence that is often incomplete and fragmented. The importance of narratives for scientific explanation has long been downplayed by philosophers of science ((Mary S. Morgan and Norton Wise 2017; Mat and Mary S. Morgan 2019), also see (Carlo R.M.A. Santagiustina 2022) in this volume), but since the second half of the 20th Century there

has been a growing awareness that scientific narratives aren't simple "just-so stories" and that they play a much more important epistemic role than assumed before (see e.g. (Goudge 1958) for one of the earliest influential repositionings of narratives in science philosophy).

This Chapter explores how narratives are used by historical scientists, and how human-centric AI systems may assist scientists in discovering and constructing more precise and testable narratives so they can achieve a deeper understanding of the world and society. More specifically, it explores the two major strategies employed by scientists: particular narratives, which aim to explain particular phenomena and events in a "standalone" fashion; and embedded narratives, which aim to identify a phenomenon or event as an instance of a recurrent pattern. It also defines a narrative as a three-layered structure that includes a fabula, plot and narration. The Chapter concludes with a first prototype for constructing narrative networks inspired by the French Revolution (1789–1799).

3.2 Particular and Embedded Narratives

Just as there exist many different kinds of narrative structures in literature, there exist different types of scientific narratives depending on the *explanandum* (i.e. the phenomenon that needs to be explained). According to (Currie 2014), there are roughly two explanatory strategies that historical scientists use, which he calls *simple* and *complex* narratives. The following two subsections provide an illustration of these two different strategies based on the French Revolution (1789–1799) (Maurois 2017; Lefebvre, Guyot, and Sagnac 1951),[1] but note that we renamed them to *particular* (instead of *complex*) and *embedded* (instead of *simple*) for reasons that we will explain in section 3.2.3, which will also explain how both strategies work.

3.2.1 The French Revolution as a Particular Narrative

In 1789, France was a large and prosperous country and one of the 18th century's world superpowers. Yet it was about to slide into one of the most important revolutions in history, whose impact is still felt today in western civilization. How could this happen?

For starters, there was something rotten in the state of France. While there was sufficient wealth in the country, the *government* was bankrupt, partly because of its expensive wars and support for the American revolution, but mostly because of exorbitant food prices and structural problems with the taxation system. In an attempt to replenish the state coffers, the French king Louis XVI made a decision he would soon come to regret: he called for a general assembly of the so-called *Estates* – the three social orders of French

[1] Both narratives reflect our understanding of the French Revolution from different sources. We underline however that we are not historians, and that scholars of the French Revolution may disagree with our narrative choices.

society of the time, consisting of the clergy (First Estate), nobility (Second Estate), and Commoners (Third Estate) – hoping that they would approve of new taxes and solve his debt crisis.

By asking for new taxes, however, the king had failed to read the proverbial room: a strong class resentment had been stirring in the country, which opposed especially the Third Estate to the first two estates. The representatives of the Third Estate, which included lawyers, local officials, and wealthy land owners, were upset about the growing inequality in French society: since each Estate had one vote, the clergy and nobility could always gang up against the commoners. This meant that the Commoners, which represented by far the biggest slice of the population, had to carry almost all of the economic burden and pay most of the taxes, while getting little political influence and social status in return.

The representatives of the Third Estate decided they would no longer put up with a system that they felt was rigged against them, and created a National Assembly on June 17, 1789 that would be *representative*: one vote per person instead of one per Estate. A clear act of rebellion, but the king was slow to react, and three days later, the representatives took the famous Tennis Court Oath – literally named after the fact that it was pledged on a real tennis court of Louis XVI – in which they swore to remain assembled until a new constitution was drafted.

The Revolution had begun... And what would follow was a period of radical changes between 1789 and 1799, depicted in a timeline in Figure 3.1. In the span of a decade, the regime of the old government (the *Ancient Régime*) would be overthrown, and a Republic would ultimately emerge after a series of violent confrontations between competing factions of the French population.

From Absolute to Constitutional Monarchy

While the first acts of rebellion were led by a relatively wealthy middle class, the general population had grievances of its own: bread prices were soaring due to crop failures and the *gabelle*, a widely hated taxation on salt.[2] By failing to feed the hungry, France was feeding the momentum of the revolution. Especially the city of Paris became a hotbed for riots, and state properties became targets of attacks and plundering. The most iconic outburst of violence happened on the 14th of July 1789, when the Revolutionaries stormed the Bastille, a medieval fortress and prison that represented in their view the monarchy's abuse of power. The fall of the Bastille went down in history as a watershed moment of the French Revolution: the point of no return.

Other major events followed rapidly. On August 4, still in the year of 1789, the National Assembly abolished feudalism, which marked the end of the three-estate system. Later

[2]The *gabelle* was abolished by the French revolutionaries in 1790, only to be reinstated by Napoleon in 1806. Still hated as an unfair taxation on the poor, it would finally be terminated in 1945 after the liberation of France at the end of the Second World War (Chazelas 1968).

that month, August 26, there was the Declaration of Human Rights, which granted human civil rights to a large portion of the population, and which became a major stepping-stone towards more equitable democracies.

A real tipping point in the transfer of power from the monarchy to the revolutionaries was the Women's March on October 5, 1789. The March started with a protest against high bread prices by women who worked at the marketplaces of Paris. A coalition of angry market vendors, revolutionary agitators and other protestors soon formed a mob of thousands of people who marched to the palace of Versailles where the king was residing. The mob forced the king to return with them to Paris and to accept the legitimacy of the National Assembly. Within a matter of a few months, the monarchy had lost almost all of its absolute power to the revolutionaries.

A Game of Thrones

Now that the Absolute Monarchy had been brought down to its knees, the Assembly tried to restore peace in the country. This became apparent at the *Fête de la Fédération* (Festival of the Federation) that was held on the 14th of July 1790, the first birthday of the storming of the Bastille: instead of glorifying the bloodshed of the year before, the festival aimed to foster national unity and even reserved a role for the king, who pledged his allegiance to the new constitution.

However, under these still waters, danger was lurking below. Different factions with literal cutthroat politics emerged that started to compete with each other for power. The first majority of the National Assembly was a coalition of moderate deputies known

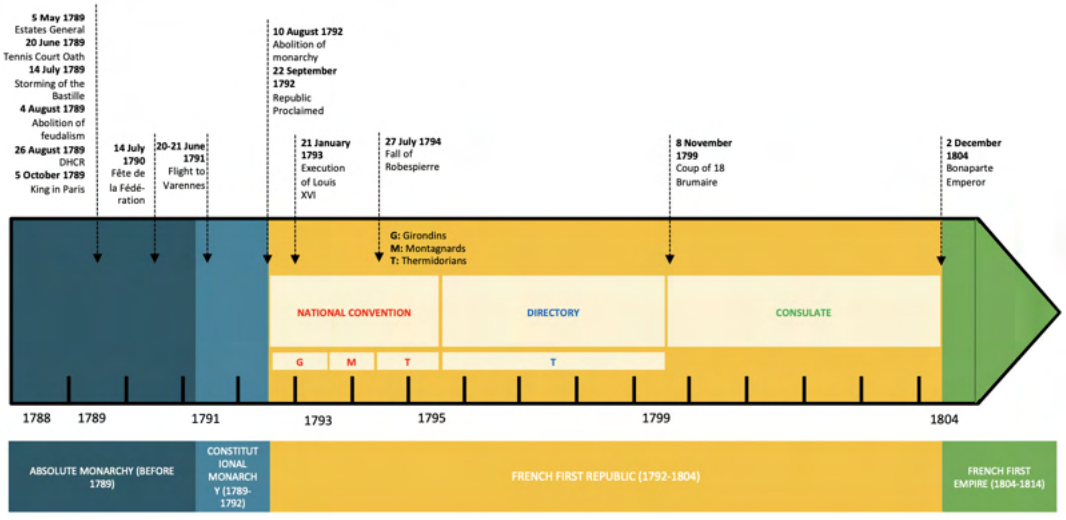


Figure 3.1: Timeline of the main events and political regimes or governments during the French Revolution (1789–1799). Translated from French and modified from: "Les temps forts de la révolution française", https://yann-bouvier.jimdofree.com/ressources/histoire/chronologie-revolution-francaise/ .

as the *Feuillants Club* who wanted to preserve a role for the king in a constitutional monarchy, but they were under attack from both reactionaries who wanted to restore the Ancien Régime as well as from a heterogeneous group of revolutionaries and anti-royalists, especially the infamous *Jacobin Club*.

Tensions quickly rose, both within and outside of France. Other European monarchies such as Austria and Prussia were getting increasingly worried that the revolution would spread to their country, and contemplated attacking France to defend the Ancien Régime or simply to take advantage of the country's inner chaos. Moreover, in June 1791, the king and his family tried to flee from Paris, presumably to start a counter-revolution. Rather than intimidating the French, these events only fuelled the flames of the revolution and public opinion radicalized even further, which made the Jacobin Club the post powerful political faction in France.

Some Jacobin deputies soon made calls for a revolutionary war against Austria and Prussia, and forced out the Feuillants from the Assembly – having hundreds of their rivals arrested and tried for treason. The ousting of the Feuillants meant that the Assembly was now more than ever prey to the power of the political clubs instead of being an autonomous governing body. However, the Jacobin Club was not a coherent group and was undergoing a power struggle itself. Especially Maximilien Robespierre rose to prominence by denouncing the war plans of the members he mockingly called the *Girondins* (named after their home department in the Southwest of France). Even though the Girondins still had the upper hand and managed to push their war through the Assembly, they would start a fierce rivalry with the dissenting Robespierrist movement, which became known as *La Montagne* (the mountain) because its members (the *Montagnards*) sat on the highest benches of the Assembly.

The war declared against Austria in April 1792 started off disastrously, however, and the Girondins were blamed for the losses suffered during the first battles. Moreover, anger had been boiling up about foreign counter-revolutionary threats and the king's failed flight attempt, which led to the insurrection of August 10, 1792, in which armed revolutionaries stormed the Tuileries palace in Paris where the king and his family were staying. In the meantime, many volunteers driven by nationalism strengthened France's army, leading to a stunning victory against invading forces on 20 September 1792. On the same day, a new Assembly called the National Convention was formed, which abolished the monarchy and proclaimed the First Republic of France.

Amidst all these events, the Montagnards seized the opportunity to take control of the country. While the Girondins hesitated about what to do with the king, the more radical Montagnards took a hard stance against Louis XVI, which gained them support from the lower class commoners (known as the *sans-cullottes*), who felt betrayed because the Girondins would not establish universal rights extending to all citizens. Accused of being royalists, the Girondins were purged from the National Convention in 1793, giving full reign to the Montagnards and its leader Robespierre.

However, as members of the bourgeoisie, the Montagnards were under pressure to satisfy

the needs of the radical sans-cullottes, which resulted in several policies enacted in favor of helping lower-class citizens and the poor, but also in extremely violent persecution of anyone who was accused of being an anti-revolutionary. Conspiracies were formed against the Montagnard's "Reign of Terror", however, which ended in a coup d'état on 9 Thermidor, year II (27 July 1794 on the French revolutionary calendar). Robespierre, the man who used to be called the "Incorruptible", was guillotined the next day. This *Thermidorian Reaction* would be marked by persecution of former Jacobins and other people who were associated with Robespierre.

The new regime, called the Directory or the Thermidorian Convention, tried to stabilize the Republic but was facing many struggles such as rebellions from royalists, former Jacobins, and the perpetual wars between revolutionaries and counter-revolutionaries across Europe. These struggles ended when a strong leader took over: general Napoleon Bonaparte. Bonaparte had led many victorious military campaigns, and ended the Directory when he came back to France through the coups of 9–10 November 1799. A new government, called the Consulate, was founded, which many historians consider to be the end of the French Revolution.

3.2.2 The French Revolution as an Embedded Narrative

Why do revolutions occur? Are there regularities or patterns in the dynamics of a revolution? One attempt at answering these questions is the influential book *The Anatomy of Revolution* by Crane Brinton (1965). First published in 1938, the book compares the outbreak and progress of revolutions to the outbreak and progress of fever, as summarized in Figure 3.2. Brinton examined and compared four revolutions: the British revolution of 1677, the American revolution (1776–1784), the French Revolution (1789–1799), and the Russian revolution of 1917.

Preliminary Stage Symptoms

Before the outbreak of the "disease", there are usually already some symptoms present. In all four case studies, Brinton noticed that the government was experiencing financial troubles, even though the society itself was prosperous. In France, the government went bankrupt not because of a lack of wealth, but because of structural problems with its tax collection system, and because attempts to reform taxes were blocked by the ruling elite. Just like the other three cases studied by Brinton, the French government was weak and inefficient.

At the same time, a prosperous middle class emerges with grievances about its socio-economic status and growing inequality. In France, these were "commoners" that included lawyers, businessmen and land owners. The Commoners were the so-called Third Estate who paid a disproportionate amount of taxes and who lacked political power because the other two estates – the clergy and nobility – would always side against them. In America

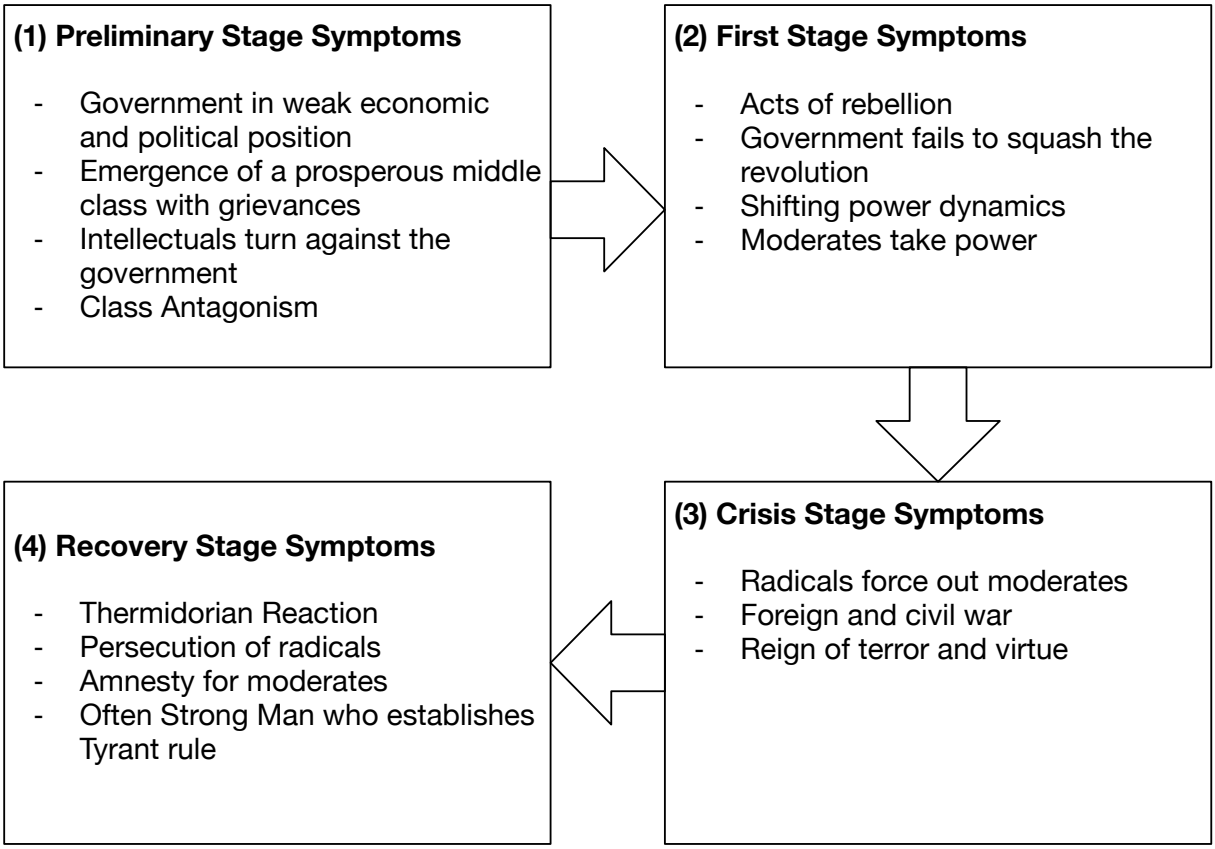


Figure 3.2: This diagram summarizes the scientific narrrative of (Brinton 1965), who compared the dynamics of a revolution to fever symptoms. All case studies are presented as narratives that are embedded in this overall structure.

these were merchants who resented the British ruling class for similar reasons. In each case, this middle class also found support from intellectuals, who increasingly turned away from the governments to demand far-reaching reforms.

First Stage Symptoms

Brinton then states that revolutions go through different stages similar to the progression of fever, with moderate symptoms at first but soon leading to a state of delirium. In the first stage, first actions are taken against an unpopular economic situation (particularly taxes), and at least two groups become opposed to each other. In France, the Third Estate of Commoners rebelled by forming a representative Assembly that would protect them from being outvoted by the First and Second Estates, which created an opposition between royalists and revolutionaries. In each case study, the government was too slow to react or failed to suppress the opposition by force, and the power dynamics shift to the new group, of which the most moderate ones take up leadership first. In the French Revolution, the Absolute Monarchy was abolished in favor of a constitutional monarchy.

Crisis Stage Symptoms

In the next stage, revolutions reach peak fever in which increasingly radical factions force out the more moderate groups through coups d'état, culminating in a reign of terror where any (perceived) opposition to the revolution is violently squashed. The Jacobins in France, the Bolsheviks in Russia, and the Sons of Liberty in America were all well-organized and disciplined but radical factions that did not hesitate to use violence to get their ways. Once in power, leaders such as Robespierre in France or Lenin in Russia prove to be authoritarians, and install a reign of terror in which all (perceived) opposition to the revolution and the new government is violently squashed. Often these swift changes in government happen against a backdrop of foreign and civil war, which pressures the governments into rapid centralization and zero-tolerance policies.

Thermidorian Reaction

According to Brinton, there is only so much that a society can take, so reigns of terror are met with a "Thermidorian Reaction", which is like a recovery from the fever. In the French Revolution, the Reaction stared with the arrest and execution of Maximilien Robespierre, which ended the Reign of Terror of the Montagnards. During this period, there is usually a prosecution of the most radical revolutionaries, while moderates receive amnesty. Often, however, such post-revolutionary societies remain unstable until a strong man such as Napoleon Bonaparte seizes control. Brinton's conclusion is that the French, English and Russian revolutions "began in hope and moderation, all reach a crisis in a reign of terror, and all end in something like a dictatorship – Cromwell, Bonaparte, Stalin" (Brinton 1965, p. 23).

3.2.3 Discussion of the Strategies

Section 3.2.1 illustrates what is arguably the most common scientific narrative in the historical sciences: the particular narrative. A particular narrative is an explanation for what the scientist identifies as a single event without necessarily trying to find out what is general or universal about the event. (Currie 2014) calls these narratives *complex* because they are typically high in *detail*: the explanation requires *specificity* and a high *diffusion* of information in order to be adequate. Indeed, one can fill a whole library section with studies that focus entirely on the events that led to the French Revolution and how the revolution unfolded over time without *embedding* these studies into a larger theory about what "revolutions" are.

Consider now section 3.2.2 on the other hand. In this narrative, the focus is not on the details or specifics of the French revolution, but rather how it illustrates a regular pattern of what constitutes a revolution (or at least those studied by (Brinton 1965)). In other words, the narrative is *embedded* in a larger framework so the specifics can be left out.

This is why (Currie 2014) calls these narratives *simple*: the information presented is less detailed, less specific and less diffuse than for the particular narrative.

We prefer the terms *particular* and *embedded* narratives because embedded narratives can become quite complex and detailed in their own right ((see e.g. Palmer and Armitage 2014, for a complex narrative that relates several revolutions to each other)). Another way to look at the two strategies is to draw analogies to *literary narratives* such as novels on the one hand (particular narratives), and *literary criticism* (that is, the comparison, analysis and evaluation of literary works) on the other (embedded narrative). Indeed, a work of fiction may contain intertextual references to other art, but must essentially be able "to stand on its own feet." Likewise, a linguist who studies the French language may refer to syntactic rules in related languages, but must ultimately provide a self-contained explanation for the grammar of French. A literary critic, then, finds motifs and themes across texts. And likewise, linguistic typologists may develop a comparative theory about the world's language structures.

If we want to develop human-centric AI systems that can assist historical scientists in constructing their narratives, we therefore need to be aware of which narrative strategy the scientist wishes to employ: a particular narrative, which offers a full and detailed explanation of a single event; or an embedded narrative, in which the focus lies on the identification of recurrent patterns.

3.3 Fabula, Plot, and Narrative

> *"The king died and then the queen died is a story. The king died, and then queen died of grief is a plot."* (Edward Morgan Forster)

So far we have used the term *narrative* in an intuitive way, but if we wish to design human-centric AI systems that make use of this concept, we need to provide a more concrete definition. This is a non-trivial task, since the western tradition of narratology goes all the way back to Ancient Greece, which has led to often conflicting views of what constitutes a narrative. We will therefore offer here some tentative definitions that serve as the foundations of our MUHAI research, but which will be further fleshed out throughout the project. More concretely, we will distinguish three concepts that together constitute a narrative:

- The *fabula* (often called the *story*) is a collection of actions, events, or facts.
- A *plot* (also called the *syuzhet*) is a structure that arranges the elements of the fabula in a causal chain or causal network.
- A *narration* (also called the *discourse* or *narrative presentation*) concerns *how* the narrative is presented.

3.3.1 The Fabula and the Dimension of Veracity

As explained in the introductory chapter of this volume (Luc Steels 2022), one of the most important dimensions in scientific narratives concerns the *veracity* of the narrative. Indeed, the primary concern of any scientific discipline is getting its facts straight. The foundational layer for every scientific narrative is therefore what is called the *fabula* in narratology, which roughly means "the story as it actually happened."

The goal of the fabula is to have a collection of facts and descriptions of events that are as objective and close to reality as possible, without trying to relate those facts with each other. For example, the Florentine *Catasto* is a historical record that offers historians raw data about the tax assessment of households in Florence and its surrounding territories between 1427 and 1429 (Herlihy and Klapisch-Zuber 1985). The Catasto includes, among others, the assets and debts of Florentine households, which helps to estimate how wealthy these households were.

In order to obtain a high degree of veracity, historical scientists must have a good estimate of the *reliability* of their evidence. For example, they must verify whether the Florentine Catasto indeed includes all or at least most of the Florentine households. When compared to historical population estimates, it turns out that the Catasto probably underreports the number of inhabitants, but that the difference is small enough to remain representative (*The Florentine Catasto of 1427* n.d.).

3.3.2 The Plot: Plausibility and Meaningfulness

The study of narratology has distinguished the fabula from the *plot* ever since its introduction by the Russian formalist school of literary criticism (Erlich 1973). The plot provides structure and coherence to the elements of the fabula through causal links, so that previously unconnected facts can now travel together as a group.[3] The simplest plot is a chain of causal links that ends with the historical event that a researcher wants to explain, though some events may be so complex that it is more appropriate to represent the plot as a *causal network*.

For example, suppose we want to explain why France celebrates its national holiday on the 14th of July. The fabula contains several facts, such as the fact that there was an attack on the Bastille on the 14th of July 1789, that there was a *Fête de la Fédération* (Festival of the Federation) on the same day in 1790, and that the 14th of July became the national holiday in 1880 – almost a century later. The plot, then, is a *plausible* causal chain that leads from the first event in 1789 all the way to the establishment of the current national holiday.

We emphasize the word *plausible* here because establishing a causal link between two events is a non-trivial task. Often the historical record consists of fragmented and incomplete pieces of evidence, so the scientist is forced to posit conjectures and fill in the gaps.

[3] The original notion of *syuzhet* conflated what we call plot and narration in this paper.

Even when there is abundant data available, causality remains tricky. A classic example is establishing the cause of death of a person, which is why physicians and medical examiners receive explicit training in verbalizing their uncertainty when signing a person's death certificate (Hanzlick 1997). The more uncertainty there is, the more hypotheses (or plotlines) may emerge in the scientific literature.

As explained in the introductory chapter of this volume (Luc Steels 2022), we consider something *meaningful* if it is *relevant* for a particular task. Constructing a plot is therefore essentially a *meaning-making* process because the historical scientist needs to decide which facts from the fabula are relevant for their scientific explanation. For example, on the 1st of July 1989, just after the start of the French Revolution but two weeks before the storming of the Bastille, the ballet *La fille mal gardée* (The Wayward's Daughter) of Jean Dauberval premiered at the Grand Théatre de Bordeaux in France (Guest 1960). Even though this piece is important for modern ballet, it is irrelevant for understanding what happened in the French revolution and is therefore meaningless in this context.[4]

3.3.3 The Narration: Structure, Style and Narrative Positioning

The third layer of a narrative concerns the *narration* or the way in which a narrative is presented to its audience. We can distinguish three dimensions in this layer: structure, style and narrative positioning.

Structure and Style of the Narration

The structure of the narration is the order in which the events of the plot are exposed. The most straightforward structure is to present the events in chronological order, in which case the structure of the narrative follows the direction of the plot. Indeed, the plot and the structure of the narrative have often been conflated, but the difference between the two have become more clear as artists have expanded their storytelling devices, such as the use of flashbacks or flashforwards in movies.

One good example is the 1994 movie *Immortal Beloved*, written and directed by Bernard Rose, which starts with the death of Ludwig Van Beethoven (played by Gary Oldman), whose testament states that all of his assets will be left to his "immortal beloved." But who is this person? The audience then accompanies Beethoven's assistant and friend Anton Schindler (played by Jeroen Krabbé), who tries to solve this mystery by visiting all of the women who played a role in Beethoven's life. During each visit, we get to see a

[4] This is not to say that a performance can never be meaningful in a historical narrative. For example, the opera *La Muette de Portici* (The Mute Girl of Portici) by Daniel Auber is often said to have played a role in creating unrest that started the Belgian revolution of 1830, though historians have downplayed its actual importance (Slatin 1979).

flashback from Beethoven's life, each time providing a piece of the puzzle until the whole plot is revealed and we figure out the identity of the maestro's immortal beloved.

Changing the narrative structure from chronological reporting to a more complex narration can be very effective for creating a compelling scientific narrative as well. One excellent example is Chapter 7 of *Life Ascending: The Ten Great Inventions of Evolution* by the evolutionary biochemist Nick Lane (2009). Early on in the Chapter, which explains the evolution of sight, Lane uses foreshadowing to whet the reader's appetite as follows:

> *"[T]he rise of molecular genetics in the last decades furnishes us with a wealth of detail, giving very particular answers to very particular questions. When these answers are all threaded together, a compelling view emerges of how the eye evolved, and from where – a surprisingly remote and green ancestor. In this chapter, we'll follow this thread to see exactly what use is half an eye, how lenses evolved, and where the light-sensitive cells of the retina came from. And in piecing together this story, we'll see that the invention of eyes really did alter the pace and flow of evolution..."* (Lane 2009, p. 175)

In the remainder of the chapter, rather than strictly adhering to the chronological order of the plot, Lane follows the chronology of the scientific discoveries that each provided pieces of the puzzle, much like a detective novel follows clues until the reader can reconstruct the causal chain of events.

But there is more to the structure of narration than simply changing the order of information. In her 1998 book, *Hamlet on the Holodeck: The Future of Narrative in Cyberspace* (1998), Digital Media professor Janet Murray predicted that the rise of interactive media would drastically impact the structure of narration because the audience is invited to play a role in creating that structure. One simple example is a website with hyperlinks: the order in which information is presented to the reader is changed whenever they click on one of the hyperlinks.

In scientific narratives as well, new forms of digital storytelling allow for the audience to co-create the structure of the narration, and in some cases even the narrative itself. One example are the CLARIAH data stories[5], which are historical narratives (e.g. about the wealth of Florence during the Renaissance) that include interactive code blocks that contain queries on a knowledge graph. Another example is the Google N-gram Viewer that allows users to type in phrases and then see how often those phrases occurred over time in a corpus of books. If the graph shows that one phrase suddenly becomes popular while another one declines, the user may become intrigued and may want to figure out what caused this shift to happen.

Whereas the structure of the narration concerns high-level choices, the *style* of the narration focuses on more concrete considerations such as register (i.e. the degree of formality of a text), verbosity, lexical diversity, grammatical complexity, and so on. Style choices

[5] https://stories.datalegend.net/

mostly depend on the goal (e.g. informing, educating, awareness-raising, promoting, and so on), modality (e.g. presentation, research paper, social media posts and blog posts, interviews, and so on) and intended audience of the narrative (e.g. children, general public, experts, policy makers, and so on).

Narrative Positioning

Stereotypically, a scientific narrative is meant to answer a question, such as how the human eye evolved. In reality, however, researchers and research groups engage in an activity that can be called *narrative positioning* (Berry 2021). Related to the concept of *framing*, narrative positioning is a technique for situating research with respect to other research endeavours using a narrative form.

The current volume of articles, especially its introductory chapter by Luc Steels (2022), is a prime example of narrative positioning by explaining how meaningful and human-centric AI compares to data-driven AI. Narrative positioning thus has a major impact on every part of a research programme, starting with which questions it is interested in answering, what kind of experiments need to be designed, which kind of technologies are needed, what kind of measures and evaluation criteria can help to track the progress of the research programme, and how results should be interpreted.

The importance and impact of narrative positioning shows that narratives not only have a major epistemic role to play in scientific explanation, but even affects and steers scientific practice itself.

3.3.4 Towards Meaningful and Human-Centric AI

Now that we have identified three important layers that make up a narrative (the fabula, plot, and narration), we can also identify how narratives can play a role in the design of meaningful and human-centric AI. In this paper, we are mainly concerned with AI systems that assist historical scientists in constructing scientific narratives in a more efficient and reliable way and to provide them with tools for better understanding complex issues in society, but the same principles could be applied to other professions as well, such as journalists who need to report on complex issues such as the COVID-19 pandemic with rapidly changing insights and a flood of information and disinformation to go through.

- *Data Veracity*: Human-centric AI systems must provide adequate knowledge representations and tools that facilitate the verification of data veracity and the assessment of reliability of pieces of evidence. This is crucial for constructing the fabula of both historical events (in which data is fragmented, incomplete and sparse) as well as contemporary events (in which knowledge is still evolving, and in which it is often difficult to distinguish facts from unfounded claims and misinformation).

- *Narrative Networks*: Meaningful AI systems must provide tools to assist humans in constructing a plot or narrative network. Examples include, but are not limited to, the extraction of causal relations from natural language texts ("understanding"), selection of relevant (and therefore meaningful) pieces of knowledge representations (e.g. through heuristic search on ontologies and knowledge graphs) for explaining a particular topic ("production"), and so on.
- *Narrative Matching*: AI systems may also assist human researchers in detecting regularities and patterns across narratives, allowing them to discover embedded narrative structures that are otherwise difficult to find because of the explosion of data and human confirmation bias.
- *Narration*: Human-centric AI systems may provide adequate tools for efficiently mediating between a human communicator and their intended audience. These may include interactive web interfaces and searchable graph visualizations ("production") and awareness-raising visualizations about information and opinion spaces ("comprehension").
- *Narrative Positioning*: A human-centric AI system must be able to explain its objectives and decisions in narrative form, including an assessment about uncertainty or decisions made based on reactive intelligence (see (see Luc Steels 2022, in this volume)).

3.4 Case Study on the French Revolution

This section presents a first prototype that explores some of the issues in the design of meaningful and human-centric AI systems that were discussed in the previous sections. More specifically, we focus on a *particular narrative* using the French Revolution as a case study, similar to the one presented in section 3.2.1. Our prototype includes the three layers of a narrative in the following ways:

1. The fabula is represented as a *knowledge graph*, taken from Wikidata in English, which is assumed to be high in veracity.
2. The plot is operationalized as a *narrative network* in which meaningful/relevant events are ordered in a chronological sequence.
3. The narration follows the sequence of the plot, and includes an interactive visualization of the narrative network.

3.4.1 The Fabula as a Knowledge Graph

The first layer of a narrative is the fabula, which can be thought of as a factbase. In our prototype, we used a *knowledge graph* (more precisely, the English Wikidata knowledge

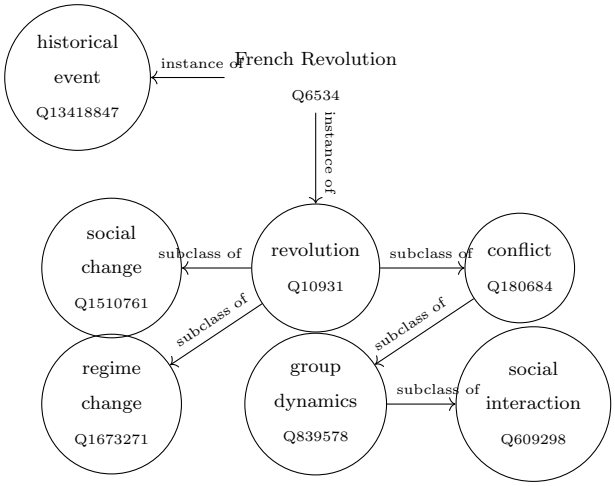


Figure 3.3: A subgraph of Wikidata centered around the search phrase "French Revolution". The nodes were manually selected to make sense of the "social and political revolution" description.

base) to represent the fabula of the French Revolution. A knowledge graph is a semantic network that may contain information about real-world entities (i.e. objects, events, situations, concepts, and so on) and relations between them. The content is encoded with the help of appropriate standardized knowledge representations, also called ontologies.

While it is impossible to have a totally objective representation of reality, we can assume that the fabula represents the ground truth and therefore scores high in the dimension of veracity (see section 3.3.1) if there is sufficient consensus about its facts. Since Wikidata is a collaborative and open endeavour, we will assume here that all of its statements are objectively true. Secondly, we assume that the fabula only contains statements that concern the four WH-questions about an event: who, what, when and where.

3.4.2 The Plot as a Narrative Network

A historical scientist who wants to construct a plot that connects all of the relevant events of the French Revolution through causal relations could in principle explore the knowledge graph manually. The French Revolution[6] is described in Wikidata as a "social and political revolution." Which facts or knowledge can corroborate this description? Figure 3.3 shows a subgraph in Wikidata centered around the search phrase "French Revolution". Exploring the graph manually from this starting point, one can see that the French Revolution is an instance of a revolution, which has interesting properties such as "social change", "regime change", or "conflict", which matches the description of the revolution as a socio-political event.

However, these links do not yet offer a narrative explanation of which events occurred during the French Revolution and how they relate to each other. The historical scientist

[6] https://www.wikidata.org/wiki/Q6534

therefore must construct a plot to make sense of what happened, starting by collecting all of the facts that are meaningful or relevant for explaining the French Revolution. Here, we manually explored which paths in the knowledge graph lead to the relevant events of the revolution.

In total we retrieved 48 events for which Wikidata often offers adequate information about the what, when and where questions (i.e. the type of events, their time and place), but little about the who question (i.e. the participants in those events). Sparsity of data is a common problem with structured data that any AI-system must anticipate, and to which we will return later in this paper. For our case study, we solved this problem of missing information by leveraging *Infoboxes* from Wikipedia.[7] A Wikipedia Infobox is a table with textual properties and attributes that summarizes the most important content about the current page. Moreover, these infoboxes also contain URL links to other Wikipedia pages. For example, if an Infobox about an event X contains a cell named "combattant" that links to a page about person A, then we can assume that there is a latent link between A and X that is missing from the knowledge graph. Using this information from Wikipedia, we were able to enrich the fabula with information about participants in the French Revolution (e.g. person A participated in event X), and to establish causal links in the plot (e.g. event X caused event Y).

Besides the selection of meaningful facts, constructing the plot of the narrative network requires the development of timelines that put events in their chronological order, and ultimately establishing a causal explanation link between events.

Figure 3.4 and Figure 3.5 display timelines that could be manually constructed from the data collected for the whole French Revolution and the French First Republic respectively. Above the blue arrow are the political regimes (first row) and governments (second row) during each period. Below the blue arrow is the timeline built with the events collected in Wikidata and Wikipedia. Events are ordered chronologically. *pre_event* and *post_event* indicate that one event was before of after another one respectively. The difference with *event_start* and *event_end* is that in the latter, one event is precisely started by another one, i.e. there are timestamps overlaps. Finally, numbered events are main events chronologically ordered identified during this period. As for the colours, a quadrille background is an event that was not retrieved originally in Wikidata, whereas a full background was retrieved in Wikidata. Dashed surroundings indicate that no corresponding Wikidata entity was found, whereas full surroundings indicate that it was.

If Figure 3.4 and Figure 3.5 permit to identify the main events during each period of the French Revolution, there are still some links missing to understand the whole picture. In particular, some links are missing or the outcomes of the events are unknown. For instance when there is a coup d'état, was it successful?

The causal line depicted in Figure 3.7 and manually built displays events together with causal links. The legends are the same than in Figures 3.4 and 3.5. A full square on the left of a node is a victory, whereas a quadrille square is a defeat.

[7] https://en.wikipedia.org/wiki/Help:Infobox

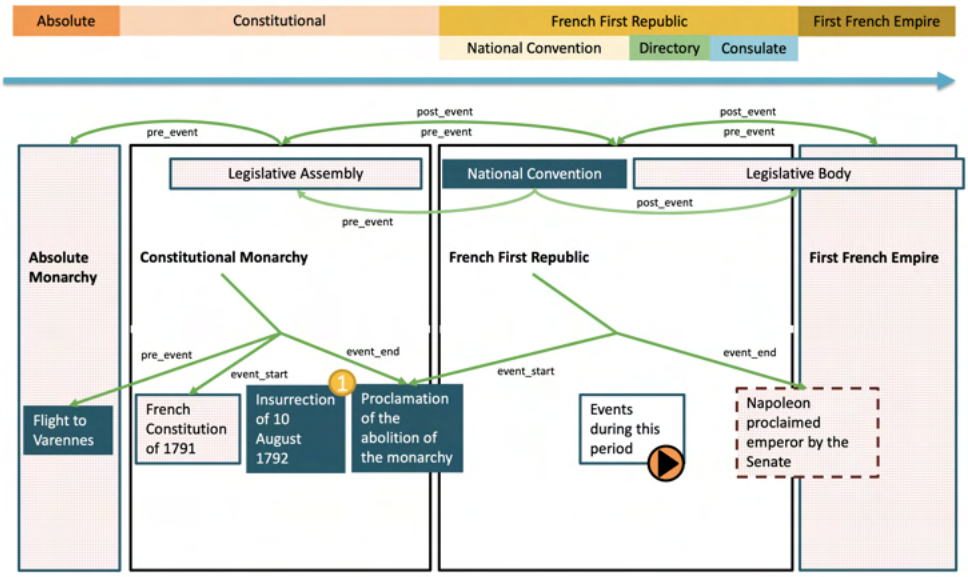


Figure 3.4: Temporal timeline extracted for the French Revolution. One can identify important events in relation to bigger events or at the intersection of other events. For instance, the event "Proclamation of the abolition of the monarchy" happens in-between the "Constitutional Monarchy" and the "French First Republic" events, and given the name of the event it is plausible that it played an important role in the transion of the two bigger events. Likewise for "Napoleon proclaimed emperor by the Senate". One can also understand how legislatures, political regimes follow each other.

3.4.3 The Narration of the Plot

Once the historical scientist has constructed a plot, they need to narrate the plot in an adequate form to their audience. For example, they can write a short essay in the style of section 3.2.1 of this paper, illustrated with timelines such as the ones of Figures 3.4, 3.5 and 3.7. They could also use novel digital tools such as an interactive web demonstration to involve the audience more intimately into the process of narrative construction, or to allow a critical reader to verify a closer inspection.

A small web demonstrator was implemented for the case study on the French Revolution. Several pages and options are available: i) collecting events about the French Revolution. The user can select paths in the graph from which to extract the events. The user can then select the type of data that should be extracted: Wikidata only or also Wikipedia, text content from Wikipedia. ii) extract infoboxes from Wikipedia. For each Wikipedia page identified for one Wikidata page, the infobox (if any) is extracted, and url links are additionally stored. Some processing options are also available. iii) build a first network. This first version would not be a finalised narrative networks since it is triples directly extracted from Wikidata, and thus do not contain events and their description (who, what, when, where). The timelines and cause lines described in Section 3.4.2 were manually built and not automatically derived from these triples. iv) visualise the network. Events were manually ordered and it is possible to slide over events over time, seeing new entities appearing over time.

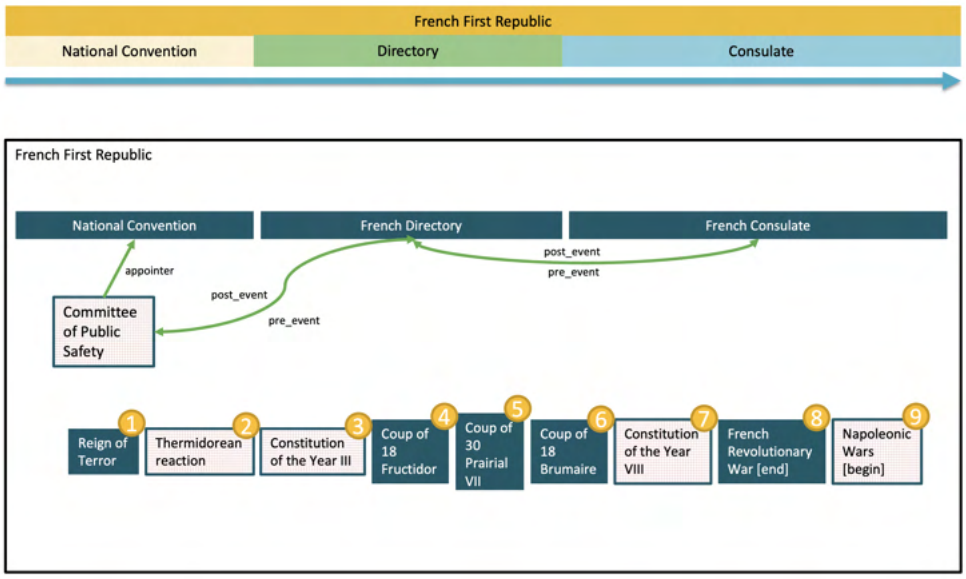
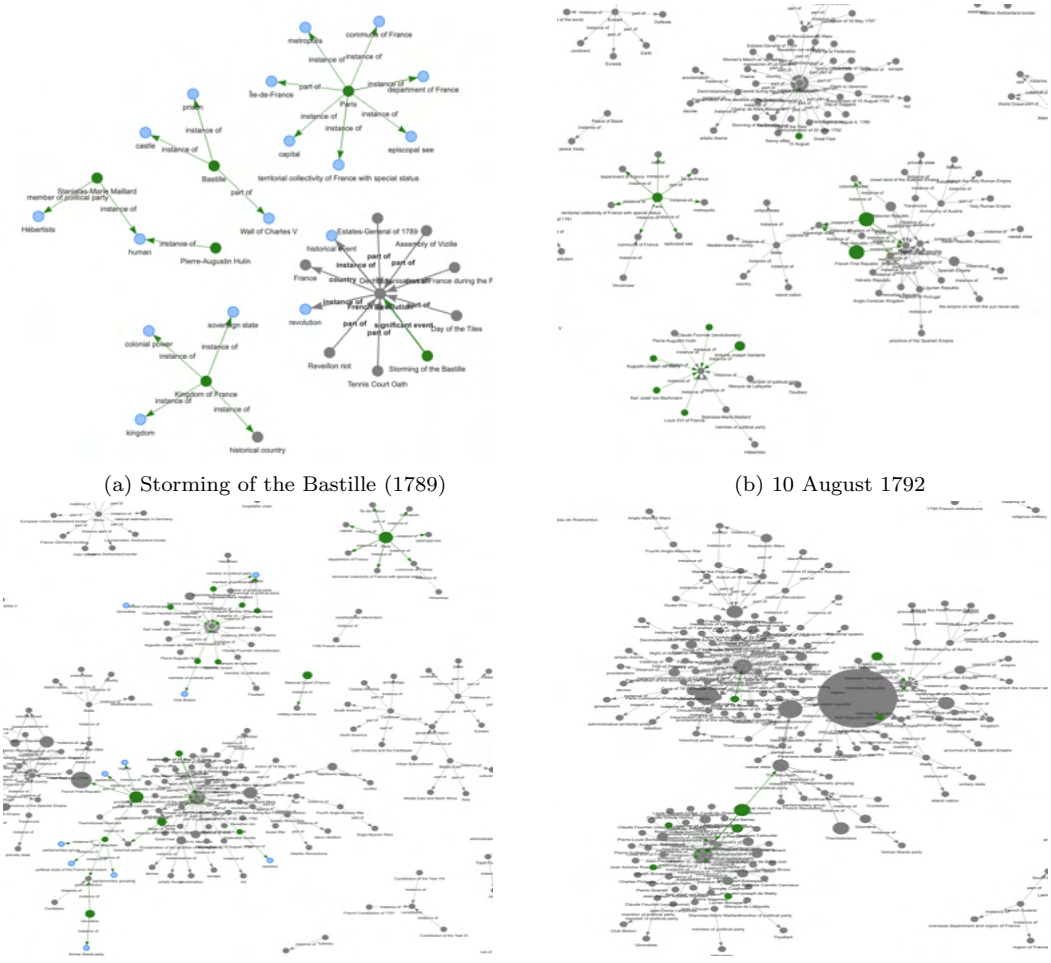


Figure 3.5: Temporal timeline extracted for the specific period of the French First Republic. One can identify the three Conventions that were described in Figure 3.2.1. One can also see the main events identified for this period.

Figure 3.6 displays a visualisation of the graph at different steps of the construction. For purposes of visualisation, the size of a node was increased if it was encountered multiple times across all infoboxes. Therefore, visually bigger nodes represent entities that are likely to have a more important role in the French Revolution. At each step for a given event, green nodes represent entities that were found in the event infobox, whereas blue nodes represent additional information. Grey nodes are nodes that are already part of the graph.

Figure 3.6a shows the entities related to the Storming of the Bastille. It is one of the first events identified during the French Revolution, where the insurgents took the Bastille as a sign of protestation of the royalist power. Different types of entities are added: the event itself, but also other types of categories like people and location (both geographical and historical). One interpretation of this experience could be that Stanislas-Marie Maillard and Pierre-Augustin Hulin both participated in the Storming of the Bastille, that happened at the Bastille prison in Paris, during the historical period of the Kingdom of France. Across the four steps displayed in Figure 3.6a, categories identified remain similar. For instance, Figure 3.6b shows the entities identified for the event of 10 August 1792. During that insurrection, Republicans were against Royalists and ended victorious, foretelling the end of the constitutional monarchy. Likewise, Figure 3.6c shows the graph built at the moment of the Insurrection of 31 May - 2 June 1793, that resulted in the fall of one political faction of the National Convention during the First French Republic, the Girondins, and the rise of another one, the Montagnards. Lastly, Figure 3.6d shows the final graph, after adding the entities from the Coup of 18 Brumaire in 1799 that brought General Napoléon Bonaparte to power, and that in the view of many ended the French



(a) Storming of the Bastille (1789)

(b) 10 August 1792

(c) Insurrection of 31 May - 2 June 1793

(d) Coup of 18 Brumaire (1799)

Figure 3.6: Graph construction at different expansion steps.

Revolution.

3.4.4 Discussion of the Case Study

In this section we presented a case study in which we emulated how a historical scientist could use present-day web technologies and knowledge bases for constructing a narrative network, including a representation of the fabula as a knowledge graph, the construction of a plot as a causal network, and a way to involve the reader in the narration through a custom-built interactive web demonstration.

This workflow has the potential to lift scientific narratives to a whole new level of scientific rigour because the research community can more reliably verify which parts of the narratives are grounded in facts (represented as the linked data of a knowledge graph) and whether those facts are high in terms of veracity and reliability; and verify which causal

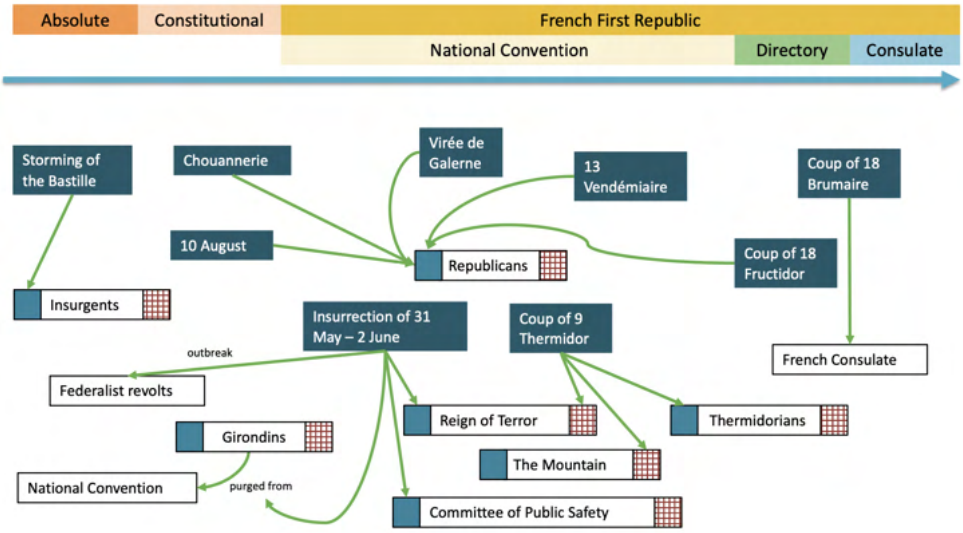


Figure 3.7: Causeline extracted for the French Revolution. The main focus is on the different political parties, and especially during the National Convention. One can see the three conventions on the causeline: the "Insurrection of 31 May - 2 June" is the transition between the Girondins and Montagnards Convention, whereas the "Coup of 9 Thermidor" is the one between the Montagnards and the Thermidorians one. The outcomes are mostly in terms of defeat or victory of politicals parties. One can also see numerous events involving the victory of Republicans: those events were indeed conflicts between Republicans and Royalists.

relations and conjectures were contributed by the scientist, and how plausible these relations are. If the fabula is undisputed, alternative plots (or scientific explanations) could be devised and grounded in the same factbase, which allows for more objective comparison of scientific hypotheses.

Of course, such a workflow would also immensely increase the burden and required skillset of the scientist to accomplish their work. Human-centric AI systems are therefore needed to assist humans in this gargantuan task. In the case of particular narratives, some important AI services could include:

- Retrieval of meaningful events from the factbase for constructing a plot (e.g. using search heuristics), to order those events chronologically, and to suggest potential causal links to the human scientist.
- Repair mechanisms to recover from data sparsity, e.g. through information-retrieval or text-to-knowledgebase parsing. Most information is still expressed as unstructured data such as natural language texts, whose volume is increasing at a much faster rate than that of structured data.
- Authoring tools that automatically ground a piece of text or a web demonstration to the fabula, and that analyze whether the narration covers the plot adequately through measures of comprehension and understanding (see in this volume for more relevant measures (Luc Steels 2022)).

Human-centered AI systems could also provide tools for helping scientists to construct embedded narratives. For example, once a narrative network of the French Revolution has been developed, an AI system could use graph-matching techniques for detecting whether that particular network can be generalized to find regular patterns such as the one presented in section 3.2.2. Not only would an AI assistant be able to find relevant matches at a faster rate than a human researcher could, it could also find generalizations that the researcher did not foresee and thereby overcome confirmation bias.

Alternatively, the human researcher could develop a general narrative pattern (such as the one for revolutions illustrated in Figure 3.2), and use an AI assistant to empirically test the adequacy of that pattern by searching for matches and mismatches in the fabula. Such applications would not only be useful for understanding the past, but also to understand present-day society. One example that is very actual at the moment of writing this article is the question how we will know when the COVID-19 pandemic has ended or at least become endemic. One piece of evidence may come from computational simulations (Lavine, Bjornsyad, and Antia 2021), while more empirical evidence could be mined through particular narratives about how past pandemic ended (e.g. the Plague, smallpox, the great influenza pandemic, SARS and others (Charters and Heitman 2021)). An AI-assistant that understands embedded narratives could help researchers to identify which biological, social, economical and political markers are worth observing, and to rapidly set up observatories that could empirically test the relevance of those markers as the pandemic unfolds over time. Human-centric AI systems could also help policy makers and journalists to better make sense of the numbers, which minimizes the danger of misinterpretation, miscommunication, and subsequent polarization about which safety measures should or should not be taken.

Finally, human-centric AI systems may also help scientists to overcome their own blind spots. As science becomes increasingly specialized, one danger is that different fields become isolated from each other so that a researcher may overlook relevant facts that are outside of their field of observation. To take an example from COVID-19 again, a virologist or epidemiologist can make recommendations for safety measures from their particular viewpoint, but lack the expertise to appreciate what the impact of those measures would be in terms of mental health, economic inequality, and so on. AI systems could help scientists to examine different plots and viewpoints to make better recommendations.

3.5 Conclusion

This paper explored the role of explanatory narratives in the historical sciences and how this could inform human-centric AI systems. More specifically, it showed that the long tradition of scientific narratives is increasingly appreciated by philosophers of science as playing an important epistemic role for understanding the world and society.

This paper then discussed two major narrative strategies – particular and embedded – and illustrated each strategy using the French Revolution. It then defined the structure

of a narrative as containing three layers: the fabula, the plot, and the narration. Through a first prototype, in which we emulated a historical scientist who wants to develop a particular narrative for explaining the French Revolution, we explored both how the notion of a narrative could be incorporated in the design of human-centric AI systems, as well as how such AI-systems may help researchers to reach new levels of scientific rigour and explanatory power in their work.

Acknowledgements

The authors wish to thank all of their MUHAI partners for their invaluable feedback on this work: the team of Luc Steels at the Venice International University, the team of Frank Van Harmelen, Annette ten Teije and Ilaria Tiddi at the Free University of Amsterdam, the team of Rainer Malaka and Robert Porzel at the University of Bremen, the team of Katrien Beuls and Paul Van Eecke at the Free University of Brussels, and Pieter Wellens at Apicbase Antwerp. We also wish to thank our colleagues Martina Galletti, Michael Anslow and David Colliaux for the many fruitful discussions; and Hiroaki Kitano (President and CEO of the Sony Computer Science Laboratories, Inc) and Vittorio Loreto (Director of the Sony Computer Science Laboratories Paris) for creating such a superb and stimulating research environment.

Chapter 4

Clinical Narratives for Causal Understanding in Medicine
Lise Stork, Ilaria Tiddi, Annette ten Teije

Abstract

In this chapter, we view clinical trials as causal narratives in the biomedical domain, as a means of understanding the causal mechanisms of treatment effects. Using Cochrane's systematic reviews as a use-case, we describe what clinical narratives are, and how they are formed, as well as tested in clinical trials. We then discuss how narratives can be represented computationally, the role of a *shared dynamic memory*, and how Artificial Intelligence (AI) techniques can be employed to support domain experts in the generation of new hypotheses.

Keywords

Causality in Medicine, Clinical trials, Drug repurposing

4.1 Introduction

Narratives serve as vehicles for understanding of past experiences, as well as hypothesising about the future. After the atomic units of an experience are segmented, grouped and interpreted in their own right, the key to *making sense* or *understanding* such experience is the narrative interpretation or *narrativisation* of the separate elements with respect to the outcome, so that the result is a coherent and understandable whole (Keven 2016). This process of narrativisation refers, first and foremost, to a form of agency leading to a

(potential) state change. An example mentioned in Chapter[1]: *"If we see someone leaning out of a window, what we see is someone who is leaning in order to do something."* In a social setting, narratives are often linked to *human* agency, motivated and justified by their goals and purposes. In the scientific domain, narratives are often of a causal nature, aiming to understand the causal mechanisms of life on earth.

By casting experiences into narratives (we call these *instantiated narratives*), and by increasingly abstracting them into more generic narratives, humans build up semantic knowledge about the world. Examples are scripts used to segment events (Anderson 2015), properties of things, or theories about causal mechanisms. Such knowledge is then continuously used to generate hypotheses about the world, by predicting and simulating future events (Schacter, Addis, and Buckner 2007). An example generic (causal) narrative from the social science domain would be that *social cohesion of a group can influence certain outcomes*, such as depression. Therefore, we can hypothesise that person X belonging to that group, is more or less likely to become depressed.

In a similar manner, the field of medicine builds upon years of experience of experimental interventions in biochemical processes and narrativisation of the possible causal effects when treating or mitigating human conditions. For this purpose, clinicians and biomedical researchers occupy themselves with building new instantiated narratives, i.e. the experimental design and validation of causal hypotheses through clinical trials. The newly acquired knowledge, gained from abstracting such narratives into more general narratives (or scientific theories), can be added to a shared scientific body of knowledge. In this chapter, we call this a *shared dynamic memory* (SDM). In turn, the SDM accumulates knowledge, from which new clinical narratives can be built. An example here would be the generic narrative that those suffering from diabetes have low blood sugar, and therefore require a glycaemic-lowering agent. Hence, we can hypothesise that population X suffering from diabetes might require a similar treatment.

Many health institutes have started to digitise as well as publish available medical knowledge on the Web as interconnected networks of data: statements that are machine readable such as *Ciprofloxacin* $\underrightarrow{isA}$ *drug* or *Ciprofloxacin* $\underrightarrow{treats}$ *bacterial infections*, in which nodes represent medical entities and edges the relationships between them. These openly available knowledge resources open op new possibilities for human-centric AI: using available medical knowledge for deliberate human-understandable reasoning in tasks such as hypothesis generation, supports human-machine cooperation, e.g., in helping domain experts come up with new, interesting clinical narratives.

In this chapter, we first describe why it is useful to view clinical questions in the biomedical domain as causal clinical narratives, how they are validated through clinical trials, and how the evidence is synthesised into systematic reviews, followed by a description of a common discovery process that leads clinicians to come up with interesting narratives (Section 4.2). In Section 4.3, using a dataset of Cochrane's semantically annotated systematic reviews as a use-case, we describe how clinical narratives can be formalised, drawing from knowledge

[1] Chapter Narratives in social neuroscience by Oscar Vilarroya

on the Web as a SDM. Finally, we describe how AI methods can use machine-readable narratives as well as the SDM to generate new hypotheses in the biomedical field.

4.2 Clinical Narratives

This section will provide background information on clinical trials as causal narratives, followed by an introduction of systematic literature reviews and their role in the biomedical domain, and a description of the common steps involved when coming up with new, clinical research questions.

4.2.1 Clinical Questions as Causal Narratives

Clinical research questions are formulated as interventional studies, set up to capture causal relationships between an intervention and a condition. In clinical trials, volunteers with a certain condition—the population—are administered an intervention with the aim of answering the specific research (sub)question, e.g. "*if a patient takes aspirin, their headache will subside*". The fundamental units here are the *human*, administering a *condition* to a *physical body*, and the patient's *bodily response* which is measured. The intervention is not limited to a real-world entity such as a chemical compound, but can also be a concept such as a change in the patient's environment or habits, imposed by a caretaker or clinician. Clinical hypotheses are not limited to drug-target interactions; other factors need to be tested including routes of drug administration, adverse drug reactions in different patients, the linkage between genetic variations, varied drug responses in different individuals, and other. A few (made up) research questions and subquestions are shown below.

Ex. 1. *Administering chemical molecule X mitigates tumor growth in patients with liver cancer.*

Sub. 1. *Administering chemical module X intravenously mitigates tumor growth in female patients between the age of 20-30, with liver cancer and back pain.*

Ex. 2. *Administering insulin lowers glucose levels in patients with diabetes.*

Sub. 1. *Administering insulin subcutaneously lowers glucose levels in children with diabetes.*

In this view, the coherent whole of the units related to a clinical hypothesis, their relevant attributes, and how they interact can be considered a *clinical narrative* (cfr. also (Hermann Kroll, Denis Nagel, and Balke 2020; Hermann Kroll, Denis Nagel, Kunz, et al. 2021)).

An example of how a clinical trial can then be cast into a coherent narrative is described using the steps enumerated below. For each step, the biomedical domain can draw from the shared dynamic memory (SDM) for segmentation, grouping, contextualisation, etc.

1. Sensory perceptions are segmented into single atomic units or observations, e.g.: `person: Jane Doe, blood sugar levels: 126 mg/dL, age: 46, drug: insulin`

2. Observations are grouped into sets of contextualised experiences, e.g.:

- `person: Jane Doe,`
- `blood sugar,`
- `levels: 126 mg/dL,`
- `age: 46,`
- `drug: insulin`

is one experience.

3. the experiences are interpreted, e.g.:

(a) `age(Jane Doe, middle aged)`

(b) `condition(Jane Doe, Diabetes)`

(c) `intervention(Jane Doe, Glycaemic-lowering agent)`

4. the interpretations are cast into a coherent narrative, e.g.:

(a) `condition(X,Diabetes)` ***`therefore`*** `intervention(X,Glycaemic-lowering agent)`

4.2.2 Systematic Reviews

When sufficiently many clinical trials have been conducted for a specific research question, researchers can conduct a *systematic review* to synthesise all the experiences obtained from the single clinical trials related to the more generic research question (Bragge 2010). Systematic reviews allow to make sense of all the evidence gathered by several clinical narratives related to the same overarching question. In other words, a systematic review can be seen as a generic narrative that clinicians generate by systematically consolidating single clinical trials (the instantiated narratives). Often, a systematic review contains a meta-analysis, i.e. a statistical analysis that includes all evidence of the clinical trials included within the systematic review. Simply because meta-analyses have more evidence to learn from, they can yield increasingly precise estimates of specificity and sensitivity.

An example of a systematic review is "*Nutritional support in hospitalised adults at nutritional risk*" (Feinberg et al. 2017), which consolidates all evidence of the clinical trials

conducted for any adult population at nutritional risk that was treated with any type of nutritional support.

Systematic reviews tend to be published online for research purposes. The Cochrane Database of Systematic Reviews is one of the largest repository of systematic reviews in health care and health policy.[2] The most well-known systematic reviews are intervention reviews, assessing the benefits and harms of interventions tested in clinical trials, but there are also other systematic reviews such as those that measure the accuracy of diagnostic tests, those that review methodologies for reporting and conducting of clinical trials, qualitative reviews, which measure aspects of interventions other than their effectiveness, and prognosis reviews which aim at systematising the most probable course of events for those with a certain disease, such as its progression.

4.2.3 Developing Clinical Research Questions

A common type of clinical research question, as discussed in the section above, is the intervention question: e.g., the measured outcome of a treatment in a given biochemical process. This *potential outcome framework* (Rubin 1974) has been used for a long time fueled by new discoveries in the lab, e.g. the discovery of a new protein. Coming up with new clinical research questions commonly includes the following steps:

Step 1. Protein-pathway discovery.
Example: *'The discovery of a new oncogene participating in a cellular pathway'*

Step 2. Drug discovery.
Example: *'The development of a chemical molecule to target the new oncogene.'*

Step 3. Clinical trial testing.
Example: *'A significant effect on tumor growth was found administering the chemical molecule to patients with liver cancer.'*

Step 4. Drug repurposing by analogy.
Example: *'The chemical molecule treats liver cancer, which resembles kidney cancer. Can the molecule treat kidney cancer too?'*

Note that the example above is not exhaustive and that we take a simplified view, where each step ingests the output of the previous step. It can perfectly well happen that a drug is found by serendipity, and thereby actually guide the protein-pathway discovery process rather than the other way around. Additionally, we purposely omitted from the example the step of literature search, usually carried in order to find gaps in an existing body of research and/or to identify bias in it (e.g. an intervention is only tested on the male population).

[2] https://www.cochranelibrary.com/about/about-cochrane-reviews

After the discovery of new knowledge for the use of a compound to treat or palliate a given disease, clinical trials are put in place to validate the effect of the compound on a target population that presents the symptoms of the targeted disease. Such clinical narratives and the systematic reviews synthesising them will contribute to validating the biomedical hypothesis and cast it into a new narrative to be added the clinicians' shared knowledge. Such cyclic process of hypothesis generation, verification by experience, narrative generation and back contribute incrementally to making sense of the biomedical knowledge.

4.3 An AI Perspective on Clinical Narratives

With the rise of digital technologies, and the capacity to store large quantities of structured data, the biomedical domain has started the shared effort of storing acquired knowledge as structured data in databases that live on the Web. In this Section, we will first introduce knowledge graphs as structures to formalise domain knowledge, then discuss a few knowledge graphs in the biomedical domain, and finally show how these can be used to build biomedical narratives.

4.3.1 Open Knowledge Graphs as SDMs

In the last two decades, research efforts have been focusing towards scaling up AI techniques to deal with the pervasive nature of the Web. The field of Knowledge Representation (KR), with its long tradition in formalising knowledge, promoted both the use of semantic technologies to easily access knowledge sources on the Web, and symbolic representations to capture knowledge. Knowledge graphs (KGs) were then born with the need of encoding real-world entities and their relationships in a machine-readable and exchangeable format (Hogan et al. 2020).

Strictly speaking, knowledge graphs are data structures describing entities and relationships in the form of triples—statements with a subject, predicate and object—which are then gathered in a directed, edge-labelled graph. Meaning and semantics are encoded using ontologies (i.e. descriptions of classes and their properties), which allow one to organise knowledge as well as perform reasoning using standard knowledge representation formalisms such as RDF, RDFS and OWL. One refers to the set of ontological classes and relationships as Terminology Box (TBox), while the statements about individuals belonging to those classes as Assertion Box (ABox). If following the semantic web standards, such as using Uniform Resource Identifiers (URIs) as unique names for things and including links to other URIs,[3] one can refer to knowledge graph(s) as Linked Data. Knowledge graphs are often built to capture either a general area of knowledge such as common-sense, or more domain specific knowledge like the medical domain. These are often manually

[3] https://www.w3.org/wiki/LinkedData

designed and less subject to errors, but also provide fewer coverage and are more costly to design. Semantic technologies allow however to integrate and manipulate data from diverse KGs, resulting in a machine-readable, large-scale shared dynamic knowledge which is not only publicly accessible but, more importantly, linked across domains, which ultimately allows machines to discover knowledge in a serendipitous way (Hartig and Özsu 2016).

4.3.2 A Shared Memory of Biomedical Knowledge

Knowledge of biochemical processes, clinical trials and other related real-world entities and concepts is carefully constructed from *shared* conceptualisations: events, agents and their characteristics, as well as compounds, side effects, outcomes, conditions, pathways, genes, etc. Below, we describe three of these resources that integrate medical terms into a single, comprehensive medical knowledge base.

The Unified Medical Language System. The Unified Medical Language System (UMLS)[4] aggregates knowledge from multiple sources, including the Gene Ontology, Drugbank, the Medical Dictionary for Regulatory Activities (MedDRA) and the Systemised Nomenclature of Medicine Clinical Terms (SNOMED CT) with the purpose of making medical terms machine-readable.

Hetionet. Hetionet[5] integrates entities and relationships from different vocabularies published on the Web into a single knowledge base describing biochemical relationships between compounds and diseases. Hetionet includes, for instance, the relationship `Compound` $\xrightarrow{\textit{binds}}$ `Gene` from BindingDB[6], or `Anatomy` $\xrightarrow{\textit{upregulates}}$ `Gene` from the TISSUES vocabulary[7]. Figure 4.1 shows the ontology in Hetionet, while Figure 4.2 shows a path on the Hetionet graph including a connection between diabetes and hypoglycaemia.

Cochrane Linked Data Project Within the Cochrane Linked Data Project,[8] experts have schematised clinical narratives into a semantic schema[9], where nodes represent the various atomic units (e.g. real-world entities such as a drug or abstract concepts such as the age of a patient) and edges represent either attributes of the nodes or narrative relations such as $\xrightarrow{\textit{intervention}}$. As mentioned earlier, a clinical trial contains a human intervention on a population with one or more conditions and symptoms, and an outcome measuring the influence of the intervention. An example of such an outcome can be

[4] https://www.nlm.nih.gov/research/umls/index.html
[5] https://het.io/
[6] https://www.bindingdb.org/bind/index.jsp
[7] https://tissues.jensenlab.org/
[8] https://linkeddata.cochrane.org/
[9] https://linkeddata.cochrane.org/pico-ontology

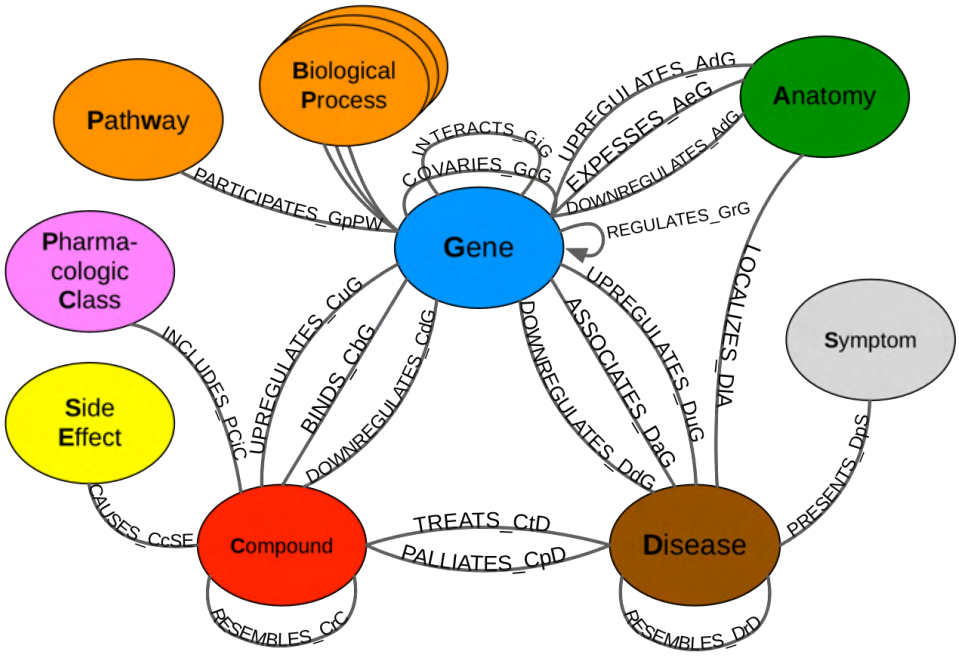


Figure 4.1: The Hetionet ontology.

Figure 4.2: Knowledge associating diabetes with hypoglycaemia, taken from Hetionet. Note that many such associations can exist. Colours for concepts correspond to those from the Hetionet Ontology in Figure 4.1.

life expectancy, or a measured decrease of glucose in the body. The schema is called the PICO ontology, referring to the Population, Intervention, Comparator (the counterfactual, which measures whether the effect would be there in the absence of the intervention as well), and Outcome of a given clinical trial. Within the Cochrane Linked Data Project, experts have annotated the collection of Cochrane's systematic reviews with concepts and relations from the PICO ontology, producing a small graph per research question (a "PICO"), in which nodes are either trial-specific, or entities from the Cochrane Linked Data Vocabulary (CLDV)[10], including types of conditions, drugs, modes for drug delivery among others.

Figure 4.3 depicts a PICO graph, i.e. the annotation of the systematic review *CD011281* (Andrade-Castellanos et al. 2016), in which the effects of delivering insulin subcutaneously

[10]https://data.cochrane.org/concepts/

or intravenously are measured in middle aged men and women with Diabetes Mellitus. In the specific, the effect on hypoglycemic events, a common side-effect of taking too much insulin, is measured.

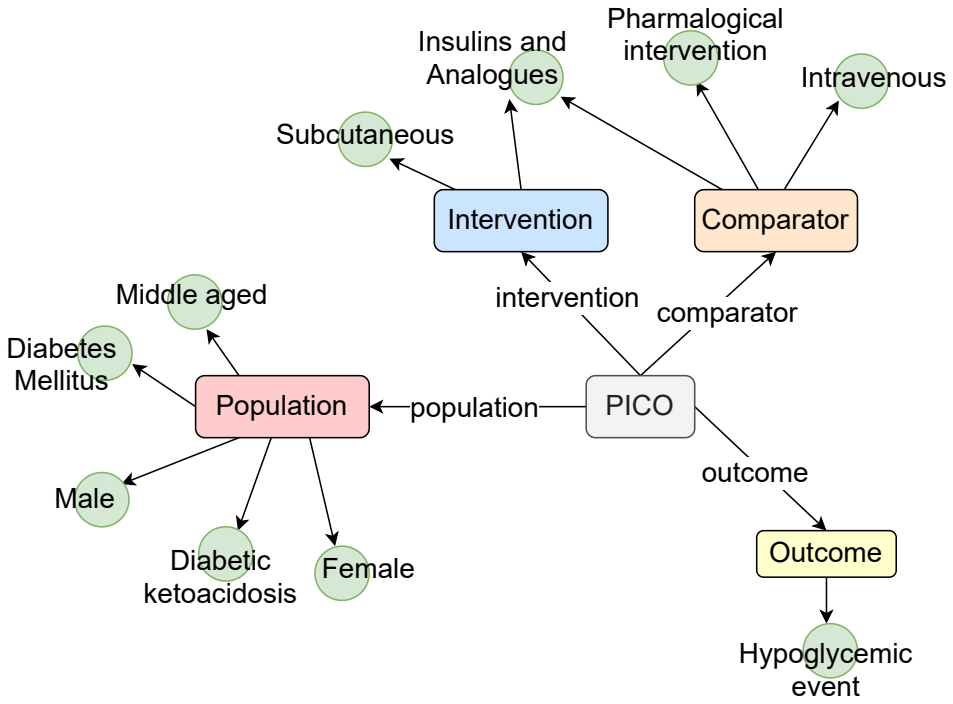


Figure 4.3: An example PICO graph, describing clinical narratives measuring the effect of (modes of) administering Insulins on a certain population with Diabetes Mellitus, on Hypoglycemic events. The graph shows all PICO elements (coloured rectangles), which link to element-specific characteristics (green circles), representing terms from the Cochrane Linked Data Vocabulary.

4.4 Clinical Narratives in Human-centric AI

It is argued that the role of AI in medicine should be purely a supporting one, and that relying on methods whose logic is nontransparent (such as the black box algorithms of machine learning) violates principles of medical ethics (Holzinger et al. 2019; Kundu 2021). Moreover, trust in such systems is hampered, something that is well-illustrated in the following example (Kundu 2021):

> [..] a conference posed the following question to its attendees: suppose you have cancer and need surgery to remove the tumor. Which of the two surgeons would you pick if you had to choose between a human surgeon, with a 15% change of dying, or a robot surgeon, with a 2% chance of dying—with the

> caveat that no one knows how the robot operates and no questions may be asked of it? All but one of the attendees preferred the human.

In this section, we show how, by employing human- and machine-readable narratives, one can enable transparent and explainable predictive systems. We do this using three task-oriented scenario's: (i) explainable link discovery for hypothesis generation, (ii) explainable graph generation for hypotheses generation and (iii) perspective or bias detection, which serves to elucidate the perspective taken in task (i) and (ii).

4.4.1 Link Discovery

Scientific discovery in the biomedical domain can greatly benefit from research into automated hypothesis generation, as finding new and interesting research questions is challenging and requires considerable background knowledge about trials, drugs, conditions and their various causal mechanisms. The task is often formulated as a link prediction task, in which a new link is predicted between a disease and an existing treatment, such as `insulin` $\xrightarrow{\textit{treats}}$ `diabetes`.

Several studies argue for the integration of a model with structured background knowledge about known cause and effect relationships within the problem domain (Blomqvist, Alirezaie, and Santini 2020), to support both the generation of hypotheses as well as their explanation (Fu et al. 2016; Drancé et al. 2021; Bakal et al. 2018; Liu et al. 2021). Rule mining ("Anytime bottom-up rule learning for knowledge graph completion" 2019; Gu and Missier n.d.) or path-search algorithms (Das et al. 2018; Ilaria Tiddi, Mathieu D'Aquin, and Enrico Motta 2014) are examples of algorithms that can discover new links in graphs, and supply these predictions with human-understandable explanations. Himmelstein et al. (Himmelstein et al. 2017) for instance, mined logical rules from a biomedical knowledge graph called Hetionet (Figure 4.1) that could explain *treats* links between diseases and treatments. An example logical rule they found was: *Compound–binds–Gene–associates–Disease*. They call such logical rules *metapaths*, given that they include not only links, but also class types of nodes. Liu et al. used a selection of ten of these rules for automated drug repurposing. They used policy-guided walks, whereby they trained a reinforcement learning agent to walk the graph, receiving a reward if the path found for a compound-disease pair matched a logical rule(Liu et al. 2021). Sosa et al. (Sosa et al. 2019) used the Global Network of Biomedical Relationships (GNBR) (Percha and Altman 2018) to develop a knowledge graph embedding-based drug repurposing method producing disease-treatment pairs. They assessed the validity of these hypotheses using a variety of sources, and, similarly to Himmelstein et al. (Himmelstein et al. 2017), discovered meaningful rules explaining newly discovered links.

4.4.2 Graph Generation

Explainable link prediction methods have proved very successful in pointing out new, interesting drug-treatment pairs, specifically in being able to focus the attention of medical practitioners to those hypotheses that are *explainable* with current knowledge on biochemical processes. While these developments are paramount in producing explainable medical AI, such hypotheses are subject to simplification. A disease and treatment cannot be seen independent of a body or population: a deviation from a 'normal' phenotype can lead to a disease, but a disease can have multiple etiologies, and one etiology can lead to different diseases. To add to the complexity, such a similar exercise can be done for diseases and their symptoms, as well as treatments and their side-effects, as each disease and treatment work together in the complicated mechanics of a body's idiosyncrasies ("The myth of generalisability in clinical research and machine learning in health care" 2020). A simple example: adults with diabetes mellitus might treat their diabetes with insulin to counter hyperglycaemia. Those suffering from hypoglycaemia, often due to a too high intake of insulin, should take care not to take any additional insulin. In those with diabetic ketoacidosis, insulin administration is essential.

The latter could be represented by not only the discovery of an interesting link, but by the generation of a small graph representing the entirety of the hypothesis: the Population, Intervention, Comparator, and Outcome (PICO, see Figure 4.3) as well as related details: age groups, symptoms, modes for drug delivery, and other. Even though explainable link prediction is a much researched topic, research into subgraph generation is scarce, and the research that exists focuses on machine-learned methods— often generative adversarial models (GANs) e.g., (De Cao and Kipf 2018)—whose predictions are not easily explained. Techniques exist that aim to explain predictions of ML models, but when applied incorrectly these still lead to wrong or misleading conclusions (Molnar et al. 2020), which is problematic especially for sensitive domains such as clinical medicine (Kundu 2021).

In this chapter, we argue for the generation of clinical graphs, representing a clinical narrative and its explanation through background knowledge (see the example in Figure 4.4). For the evaluation of clinical narratives, we formulate four metrics that are in line with the various dimensions of understanding discussed in Chapter 1. The generated narratives should aim to maximise the following metrics:

1. **Coherent**. Here, we refer to schema-correctness. Inconsistencies in the use of a semantic schema should be limited, or avoided altogether. For example, a hypothesis generated using a triple such as a given population $\underrightarrow{treats}$ a given outcome is incorrect according to the original schema, and therefore should invalidate the generated hypothesis that includes such triple.

2. **Integrated**. With a narrative's integratedness, we refer to how interconnected the narrative's entities are through background knowledge, facts such as disease-gene associations from the SDM. Let us take the example of testing the effect of insulin on hyperglycaemia for a patient with diabetes. With a knowledge base of biomedical

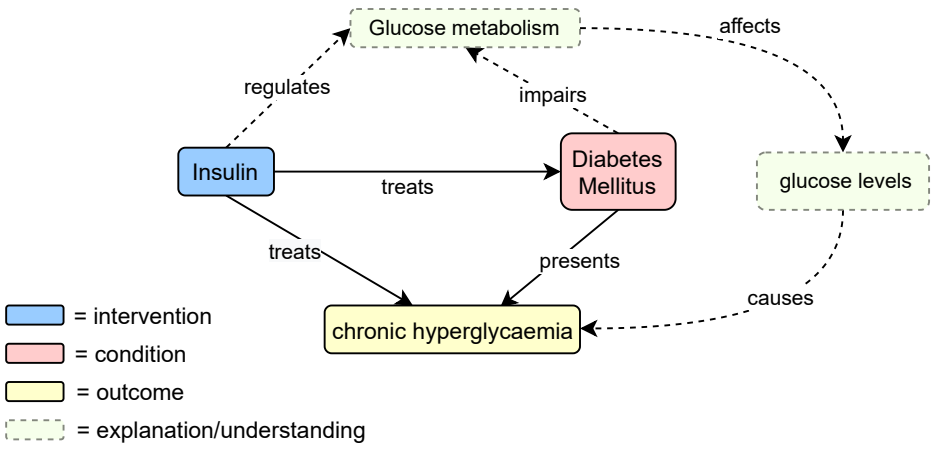


Figure 4.4: Metrics explained: solid boxes indicate a simplified PICO graph, and dotted lines indicate reasoning through background knowledge. The effect of treating a population of diabetics with insulin, measured on hyperglycaemia. The hormone `insulin` promotes glucose metabolism: the uptake of `glucose` from the blood. When there is too little `insulin` to regulate the process, which is the case for those with `diabetes`, taking exogenous `insulin` can decrease `glucose` levels to combat `hyperglycaemia`.

knowledge, one can discover causal associations between insulin, hyperglycaemia as well as diabetes, which are all tied to `glucose` levels, see the simplified PICO in Figure 4.4.

3. **Compatible**. A narrative's compatibility refers to its compatibility with knowledge from the SDM. Examples are learned logical rules such as the two example rules shown below, based on Figure 4.4, in which X and Y refer to variables in the head atom of the rule, and V_i to variables in the body atoms).

$$treats(X,Y) \leftarrow regulates(X,V_0), impairs^{-1}(V_0,Y) \tag{4.1}$$

$$\begin{aligned} treats(X,Y) \leftarrow\ & regulates(X,V_0), affects(V_0,V_1), \\ & causes(V_1,V_2), presents^{-1}(V_2,Y) \end{aligned} \tag{4.2}$$

4. **Relevant**. In order to discover whether a clinical narrative is relevant, it should be relevant to domain experts. Therefore, generated narratives should be assessed by clinical experts: instantiated narratives, the integrated background knowledge, as well as the generic narratives that help produce narratives.

We can additionally say that background knowledge from the SDM can explain the meaning of a narrative on various levels of granularity; see the green boxes and dotted lines in Figure 4.4 as well as the logical rules listed above: the two rules as well as their instantiated paths in Figure 4.4 can be viewed as examples of such levels, Rule 2 being more fine-grained than Rule 1.

4.4.3 Perspective/Bias Detection

A third relevant task to mention is the detection of perspective when using background knowledge in the generation and explanation of narratives. This perspective or *bias* greatly affects the reasoning of human-understandable AI systems relying on such knowledge. For example, some knowledge bases might focus on knowledge related to psychological mechanisms, whereas others rely on biological pathways. When reasoning systems depend on one or more types of knowledge for their explanations, we can say the system takes a certain perspective for a task such as hypothesis generation. Predictive systems should aim at making such a bias transparent to the user, for instance by using algorithms to detect a system's perspective (Ilaria Tiddi, d'Aquin, and Enrico Motta 2014; Soulet et al. 2018). When a system has knowledge about genes and pathways, but not about differences in anatomy between men and women, it would not be able to explain the reason for only testing the treatment against menstrual pains on a population of women.

4.5 Conclusions

Narratives are fundamental means to understand past experiences and hypothesise about the future. Narrativisation is the process of segmenting, grouping and giving meaning the single units of an experience, so that the result is a coherent and understandable whole. In this chapter, we introduced the idea that clinical trials serve as causal narratives for the understanding of causal mechanisms in the biomedical domain. Using Cochrane's Systematic Reviews as a use-case, we described first what clinical narratives are, and how they are formulated by clinicians, fueled by, amongst others, new discoveries in the lab. We then presented state-of-the-art AI techniques for explainable hypothesis generation: discovering new disease-treatment links through available background knowledge. We argued that, even though these techniques have furthered the field of explainable hypothesis generation, viewing hypotheses as single triples possibly reduces the ability to capture and thereby understand intricate dependencies. We proposed an extension of these models from link prediction to a more fine-grained representation of clinical hypotheses as subgraphs, and propose metrics to maximise as well as evaluate their meaningfulness. Lastly, we emphasised the importance of elucidating perspective or bias, which a predictive model takes when relying on a certain source of background information for hypothesis generation.

.

Chapter 5

Narratives in social neuroscience
Oscar Vilarroya

Abstract

Narratives are a basic component of human cognition. They are the tool the brain uses to make sense of our experiences, and to build our knowledge about the world and about ourselves. Yet, neuroscientists have been reluctant to use the notion of narratives in their theoretical frameworks and experimental paradigms(Willems, Nastase, and Milivojevic 2020). In this article, I will present some studies about the area of moral values that show that the notion of narratives must take a central role in the future of social neuroscience.

Keywords

Social Neuroscience, narratives, sacred values.

5.1 Introduction

Humans are a narrative species. We employ narratives as a way for understanding our experiences, make sense of ourselves and the world, as well as to communicate with our fellow humans. In short, the narrative faculty is a basic tool of our mental life. We tell ourselves things constantly, involuntarily and inevitably, telling ourselves what is happening, from the simplest to the most complex thing. If we see someone leaning out of a window, what we see is someone who is leaning in order to do something, whether it be to pass the time, snoop around, or to get some fresh air. If we see a car pass by, we see a car driven by someone that is going somewhere for some reason. If we smell an omelet, we imagine someone cooking it for someone to eat. If we hear someone speaking, we interpret it is as part of a conversation between two or more people. Not even an apple on a table escapes our narrative drive: we don't perceive the apple and that's that, rather we perceive it as part of a context about the apple, where it is and the person that put it

there, and we incorporate this into our understanding of the situation.

In addition, telling ourselves things is an involuntary act, like breathing. It's not something that we deliberately decide to do, in spite of our being aware of it. It's true that we tell ourselves things when we explain to someone what happened to us, as is also true that telling things is what writers, filmmakers or playwrights do professionally. But the telling of things that I'm talking about is not a deliberate thing, but an involuntary one. It's an activity that our brain cannot stop doing, as it is integrated in our perceiving and understanding the world. All our mental machinery uses narratives. The brain is programmed to tell everything that happens to us. In the same way that the retina or the ear cannot stop registering and processing data in a particular way, our brain has evolved to tell itself everything that happens to us. Everything, absolutely everything that happens to us has to be told. From the moment we open our eyes in the morning, until we close them at night, and, even more, when we dream. We wake up, check the time, and then immediately think "the night passed by too quickly", or "we're more tired than we should be because we worked too much the day before", or "the day is darker than it should be".

Extending the 'narrative paradigm' that Walter Fisher (Fisher 1987) proposed for human communication, we could say that human cognition strongly depends on the use of narratives (see table 1)

Narrative paradigm	Rational paradigm
1. Humans are narrators	1. Humans are rational beings
2. Decision-making is based on "good reasons"	2. Based on "good arguments".
3. Good reasons are based on historical, biographical, cultural or character aspects.	3. Arguments are based on logical criteria.
4. Rationality is based on the sense of internal consistency and reference to past experiences.	4. Rationality is based on the veracity and the formalities of the processes of reasoning.
5. We experience the world as narrative.	5. The world is understood as a set of logical relations that are identified via reasoning.

5.2 The adaptive function of narratives

Evolution has led humans to become a narrative species, that is, a species that tells itself what happens to itself and what happens around them via narratives. The first narratives that we created were prelinguistic. Later on, with the development of modern languages, we were able to create stories that became more and more sophisticated. We acquired that ability to include multiple persons, things, or multiple causes, to relate some narratives

to others to create a complex narrative structure, as well as to modify the minimum narrative structure itself, as poets, among others, often do.

This evolution towards our narrative nature was not binding. In the same way that we could have adopted a minimal narrative structure different from the one we adopted, we could have continued to evolve without having become narrative beings. To begin with, there doesn't seem to be any trace of our explanatory drive in species very close to us, such as chimpanzees, gorillas or orangutans, which have been able to reach the present without the need to explain the world. In fact, these species do not seem to need explanations in order to properly manage their relationship with their surroundings. A primate can learn to associate a danger sign with a real danger without having to explain anything. Knowing your surroundings and acting intelligently does not require any narrative faculty.

Therefore, it is reasonable to conceive the possibility that evolution could have led us down a path in which we did not seek explanations for what happens to us. We could have followed an evolution which was more or less parallel to that of our closest relatives, the chimpanzees or the bonobos and continue to manage our relationship with our surroundings and with our fellow human beings without having to construct narratives. But the fact is that this was not the case.

The innovation that narratives introduced took hold and was sustained because it produced very important adaptive advantages. And that is also key. If it had not produced adaptive advantages, telling stories would not have been incorporated into our species as a universal trait. It is a very costly activity that uses a lot of mental resources for a long time. Therefore, our predecessors had to divert these resources from other vital activities and that diversion had to provide some advantage for it to remain part of our way of being and doing.

But what is the adaptive function of narratives? The role of narratives in understanding what happens to us and around us seems to suggest that the function the story fulfilled was to reliably represent what happened around them. However, the modern study of narrative has revealed that the function of the story is not that of reliably representing what happened, but of giving meaning to what happened.

Therefore, we should understand narrative as the way we humans process experience with the goal of establishing its meaning. In this sense, a certain narrative competence is what helps us to make sense of the experience we live, and to connect it with previous and future experiences. Narratives, in this approach, are intrinsically experiential involving cognitive, emotional, embodied and situational dimensions.

The idea of the story as a "meaning-maker of life experiences" is not new. The value of the narrative in giving meaning to our experiences has always been recognized. The strongest defender of this idea over the last fifty years was the psychologist, Jerome Bruner. Brunner sustained that we pass through life telling stories with the aim of giving meaning, coherence and continuity to our experiences. Following Bruner, we could say, in a very concise manner, that giving meaning comprises integrating what happens in function of our present and past situation, our motivations and desires, and the context

in which what happens, happens. For each story that we draft, this includes our mental and physical state, our plans, needs and expectations of what's going to happen (and also what we don't want to happen), although, likewise, it should include much more general aspects, like our personality, our experience, our values and political ideas, among many other things. The narrative faculty is not, therefore, a representation of the world detached from the person who formulated it, rather they must be fully integrated into it, with their nature, their way of viewing the world and their past experience. Let's put an illustrative case, like a football match. Two fans from rival teams that are playing a match could explain to themselves the same game play in very distinct ways, and the two could describe what happened in a very different way, although both could do so in the most honest and sincere manner possible. They're not lying, rather they have told themselves what happened in accordance with their wishes, motivations and expectations.

5.3 Narratives in social cognition

Narratives have a very important role in the social dimension of human social life. Our social life depends on a class of psychological components that can be subsumed under the term of "values", that is, representations of desirable situations and/or behaviors that are motivated by a narrative view of the world (Haidt 2008). It is the narrative explanation of our social life that provides meaning to our social attitudes and provides justification for past and future social actions. Values are important because they determine the way we take decisions in our social life. Classical decision-making theories generally assume that individuals choose among available options, weighing up the potential costs and benefits of each option as well as their likelihoods. However, such theories may be inadequate to explain or predict social behavior, because such behavior is shaped by values.

5.3.1 Sacred values

Let's take the example of one type of values, "sacred values" (Atran and Axelrod 2008), (Baron and Spranca 1997). Sacred values (SVs) are key components of social and cultural cognition, in that they constitute beliefs that define our primary reference groups. In this sense, SVs are the cornerstones of belief systems that incorporate core values (e.g., religious commitment, family ties, group honor, justice, and patriotism). SVs clearly are not just specific ideas that cognizers hold dear. They are ideals that people are willing to fight for and defend with their lives if need be. SVs are the driving force behind in-group versus out-group actions, regardless of risks or expected outcomes. They also elicit feelings of outrage when trespassed.

Empirical studies in multiple cultures and hotspots across the world indicate that sincere attachment to SVs entails:

1. Commitment to a rule-bound logic of moral appropriateness to do what is morally

right no matter the likely risks or rewards, rather than following a utilitarian calculus of costs and consequences (Atran 2003),(Ginges et al. 2011),

2. Immunity to material trade-offs, coupled with a "backfire effect", where offers of incentives or disincentives to give up SVs heighten refusal to compromise or negotiate (Dehghani et al. 2010), (Ginges 2007),

3. Resistance to social influence and exit strategies (Atran and Henrich 2010), (Sheikh et al. 2012), which leads to unyielding social solidarity, and binds genetic strangers to voluntarily sacrifice for one another,

4. Insensitivity to spatial and temporal discounting, where considerations of distant places and people, and even far past and future events, associated with SVs significantly outweigh concerns with here and now (Atran 2010),

5. Brain-imaging patterns consistent with processing obligatory rules rather than weighing costs and benefits, and with processing perceived violations of such rules as emotionally agitating and resistant to social influence (Pretus, Hamid, Sheikh, Ginges, et al. 2018).

The notion of sacred values has accumulated a compelling corpus of findings and has been studied in the context of deep-seated political conflicts (Atran 2021). Field studies and experiments with populations involved in armed conflict find that self- reporting of support for violence appears insensitive to material costs and benefits, and asking people to trade sacred values for material benefits provokes moral outrage. In recent years, this concept has been revisited and enriched by historical and anthropological analyses in combination with a series of experiments carried out in the context of deep-seated political conflicts, such as the Israeli-Palestinian conflict, the Iranian nuclear program, the Muslim-Hindu conflict, as well as in other forms of cultural conflict (Atran 2021). Sacred values have proven to be behaviorally different from non- sacred values in that people holding them show resistance to instrumental trade-offs, even when the offer is increased to an indefinitely large amount.

5.3.2 The neurocognitive basis of sacred values

Neuroimaging findings carried out by our group support the notion of sacred values, revealing a differential neural activity between sacred values in contrast to (culturally relevant) non-sacred values. We found that sacred value choices involved less activation of brain regions previously associated with cognitive control and cost-benefit calculations (Hamid et al. 2019), (Pretus, Hamid, Sheikh, Gomez, et al. 2019), (Pretus, Hamid, Sheikh, Ginges, et al. 2018). Specifically, our studies point to a network of brain areas, including the dlPFC, IFG and parietal cortex, as key regions underlying brain differences between sacred and non-sacred values with regard to decisions about making costly sacrifices, including fighting and dying. At a level of main effects, the bilateral dlPFC, a neural hub for

evaluative cognitive processing, is less involved during sacred versus non-sacred value decisions, suggesting a decreased reliance on cognitive control functions during sacred value choices. Overall, these observations are consistent with the idea that choices involving sacred values are less dependent on cost-benefit calculations than choices involving non-sacred values, and the view of sacred values as moral imperatives guiding goal-oriented actions. They are also consistent with the role of utilitarian thinking in moral cognition, and with models that distinguish between more cognitively deliberate versus more affectively-driven reasoning. In addition, they dovetail with behavioral data showing that material incentives and disincentives, which can otherwise successfully bias utilitarian reasoning based on cost-benefit calculation, are less effective in influencing behavior when sacred values are at stake.

In addition, we also found that sacred-values' attitudes were influenced by peers opinions. Community feedback shifted willingness to fight and die for sacred values in the direction established by peers. Interestingly, change in judgment predicted neural activity in the dlPFC, which provides evidence that individuals who most changed their willingness to fight and die ratings for sacred values after the social manipulation also recruited neural areas associated with cost-benefit processing to a greater extent during the feedback paradigm. Our findings suggest that even when social network interventions are unlikely to reduce commitment to a sacred value, they could reduce adherence to violent options.

We have also found, the combined evidence between the behavioral and neural responses to both sacred value processing and willingness to fight and die indicates that social exclusion in young and vulnerable individuals may increase similarities between non-sacred values and sacred values in terms of heightened left inferior frontal activity and greater expressed willingness to fight and die. The findings point to social exclusion as a possible contributing factor to radicalization, in line with analyses from political science and criminology (Pickering, McCulloch, and Wright-Neville 2008). If so, then counteracting social exclusion and sacralization of values should be considered in any intervention or policy aimed at preventing radicalization.

5.4 Negative consequences of social narratives

Human communication has evolved in the 21st century towards new forms of representing, sharing and consuming news. The digital technology advances have made it much easier to produce and spread information in all types of formats. Today, any user, with just a computer and some computer programs, has the capacity to become their own information broadcasting channel, with the possibility of reaching all four corners of the earth, immediately. The technological breach between the large information broadcasting media and the citizen with a computer and internet connection is becoming smaller and smaller. The new digital ecosystem is overpopulated with content creators and diffusion channels.

This technological revolution has, obviously, its negative consequences, like the malicious

creation and diffusion of information by people and groups with commercial, ideological or personal interests. One of the most notorious phenomena of this trend is what is known as fake news. Fake news is a misleading narrative that is timely, spectacular and belligerent in favor or against someone, a group or an idea and that is spread by internet in order to harm this person, group or idea as much as possible. The technical ease of creating and spreading digital content, the absence of technical or legal impediments for the malicious use of these contents, and the uselessness of traditional filters to ensure the quality of the information have caused fake news to have a great media and social impact.

Another case is the phenomena known as "narrative bubbles" or "echo chambers" as they're also known. These phenomena occur under special circumstances, where a group of people convince themselves, in a short space of time, of a set of narratives that create a full, exclusive and militant vision of something happening to them or happening in their surroundings, and which leads them to carry out extraordinary collective actions to impose their narrative. The truth is that a large part of the population lacks the tools to evaluate the veracity of fake news or echo chambers, even the most flagrant misleading narratives. This is made even worse in the case of younger people given that they have consumed most information directly on internet and digital platforms, and their interaction with this medium has been direct and completely based on trust. Unfortunately, the fact is that the evolution of new technologies has, as yet, not led to the establishment of some criteria to use as a base for evaluating narratives, which are the reliable media and which not, or what are the usual mechanisms for deceit. And this has had undesirable consequences in the past, and now has even more, and there will be many more in the future.

5.4.1 Factors influencing sharing misleading narratives

There are no personalities or levels of intelligence or of cultural education that are preferentially associated with consumers of misleading narratives (Douglas, Sutton, and Cichocka 2017). It's true that the greater the educational level, the more tools available to avoid or combat the misleading news. But what we have to understand is that, the most important thing that determines whether we'll be taken in by fake news is the motivation to accept what the story wants or affirms. One may be very intelligent and cultured, but if we want what the news affirms to be true with all our heart, we will lower all our intellectual defenses, inhibit all reflection on possible risks or contrary arguments to that desired and will enthusiastically assume the fake news.

Conspirative Predisposition. By conspirative disposition we mean the tendency to accept explanations where the cause of an event is attributed to persons or groups of persons with power, whose participation in this event they intend to cover up. The belief in conspiracies is as old as the existence of conspiracies, that is, since humans were a social species. Nevertheless, the conspirative disposition does not refer to believing that conspiracies exist, rather in believing that almost everything that occurs conceals a conspiracy. It's this exaggerated predisposition that facilitates the gullibility to fake news.

The factor that best predicts whether someone has a predisposition to conspiracies is whether they have previously exaggeratedly believed in conspiracies, which is not very useful. There are, however, some features that have been associated with people that have a conspirative disposition. In first place, the conspirative disposition has been related to the perception of a lack of control over their own lives. Psychologists have developed the concept of perception of control, which has other possible denominations like, among others, control locus, personal autonomy or auto-efficacy, to refer to the way in which a person perceives the power they exert over their own lives. At one extreme of the perception of control, there are people that perceive that they can manage and do manage their own lives. At the other extreme, there are people that feel powerless, that believe that their lives are determined by external forces, be it luck, fate or other people. Psychologists have demonstrated that people with a sensation of control over their lives have a better psychological well-being, and even better health, that those that don't have a sensation of control. Lack of control leads to depression, stress and anxiety disorders and, curiously, to particular thought tendencies, like the conspirative thought, as the person extends the lack of control not just to their own lives, but to the events that surround them.

The second trait pertaining to people with conspirative disposition is their tendency towards magical thinking, that is, the tendency to more easily employ supernatural causes in their explanations. As we saw in the chapter "The narrator's reality", magical thinking is a style of thinking that humans have recurred to since we began to tell ourselves what was happening to us. History is full of cultures with magical reasoning that tries to explain inexplicable natural and historical phenomena. The human narrative drive impels us to obtain an explanation, no matter what, and if we can't find a reasonable cause that explains it, then we invent a magical one in order to end the story. However, there are people who tend towards supernatural explanations much morefrequently than other people, and faced with a prosaic or a magical explanation, they prefer magical.

The last trait identified among persons with a disposition for conspiracies is the tendency to what we could call epic thinking, that is, the tendency to prefer particularly melodramatic stories in the explanation of important events, especially those that refer to social life. The epic tale is a hyperbolic story, with exaggerated attitudes, based on the struggle between the forces of good and evil, with heroic or villainous characters, and full of extreme values of goodness and badness. These explanations are very usual in historical-populist stories and align very well with situations where a group or an idea is in danger.

All together, these three factors, lack of control, magical thinking and epic thinking are largely found in those people with a tendency to believe in conspiracies. It's true that these three traits are also seen in some psychopathologies, but it's important to highlight that a conspirative disposition is not synonymous with psychopathology. Even though lack of control, magical or epic thinking are seen in some psychiatric disorders, such as paranoid personality disorder or schizophrenia, people with a conspirative disposition normally have a completely normal life, adapted to their environment, which is not characteristic of psychopathologies. What's more, people with a conspirative disposition do not create

their own theories, rather they assume some of the ones that have reached them. Finally, people with conspirative tendencies tend to seek and believe in conspiracies that support their prior beliefs or the group that they identify with.

Biased prior knowledge. Partisan knowledge of a subject is, nevertheless, the most important factor that determines gullibility to fake news. The availability of access to any type of information that internet provides gives anyone the opportunity to get information on any subject. The problem is that false, biased or incomplete information is easily available, and anyone that sustains a partisan story concerning a subject can easily find all the information they want that supports it. The studies are clear, the greater the intensity of partisanship involved in the story, the greater the probability of gullibility for the information that supports their story. The idea of partisan knowledge as the cause of social behavior is not new to science. Researchers in political, sociological and communication sciences, some time ago, identified sectors of the population that uphold a partisan knowledge of a subject, and that refuse to modify it with evidence that contradicts it. In the case of the witches of Salem, for example, the conviction of the existence of witches became a dogma of faith that determined the way in which the community interpreted the behavior of the girls.

The authorities know from experience that a follower of a group is particularly refractory to any campaign to modify their partisan knowledge. The more intense the attitude they feel towards the story they uphold, the more information coherent to this story they consume, and the more they reject the information that refutes it. This leads to a collective attitude that denies any dissonance of their partisan knowledge. In addition, partisan followers habitually mistrust governmental or corporative sources of information. Their predisposition to believe in conspiracies of the political and/or economic powers makes enemies of any public or private official institution that approaches them. In the case of vaccines, for example, the anti-vaccination groups are convinced that the pharmaceutical industry has the aim of selling vaccines even though they know of their unwanted effects, and that they have infiltrated government sectors to defend their interests, hence no information coming from public health institutions is trustworthy.

5.4.2 Modulators of sharing misinformation

Emotional modulators. Experimental studies on misinformation show that the way we consume fake news depends on the emotional state of the consumer, and especially on two types of emotions: anger and fear. They showed that these two emotions determine, independent of each other, the acceptance of fake news.

Anger induces the partisan evaluation of the information, such that false information coherent with the position they uphold is more easily assumed. In contrast, fear doesn't act via partisanship. The person that experiments fear wants to stop being afraid and this predisposes them to accept information that reduces their fear, independent of how true the information is. What happens is, for a partisan follower, everything that is coherent

with their partisan story usually reduces their fear.

Cognitive modulators. There are various cognitive mechanisms that facilitate gullibility to misleading narratives. One of the best-known ones is repetition. Repetition is a classic strategy in campaigns of persuasion. It's been known for some time that, the more a message is repeated, the more possibilities of it being considered true. In fact, just one repetition is sufficient for it to start to have an effect. It has been shown that, even when we have prior knowledge to evaluate the falseness of a message, the mere repetition of the false message induces the sensation of the irrelevance of our own knowledge. In other words, we know that something is not true, but repetition of the lie makes us ignore that we know it's not true.

In the *online* digital ecosystem, repetition seems to work the same as, or better than, traditional media. Recent studies, using Facebook as the online platform, have proven that repetition acts as an inducer of gullibility, independent of how lacking in credibility the news is ("The Earth is flat"), or even when accompanied by contrary information based on "fact verifiers", that is, internet agents whose function is the verification of the facts spread on social networks.

Why does repetition turn us into gullible people? Repetition is supposed to induce gullibility because it makes the message familiar, and familiarity is a potent factor of credibility. This is one of the cognitive biases that we saw in the chapter "The rational narrator, or almost". Also, familiarity helps in the acceptance of a message, as it's easier to process a familiar message than a new one, and this makes the familiar one more acceptable.

Another cognitive modulator is what researchers have baptized as the "partisan boomerang effect": If someone is very sure about their partisan story, the presentation of irrefutable data against the partisan story not only doesn't make the receptor of this data doubt their story, rather they become even more convinced of it. This has been shown in numerous studies. For example, in a study on people that supported the Iraq invasion based on the presence of weapons of mass destruction, the researchers presented irrefutable data as to their inexistence. Despite this, the participants came out of the experiment believing more firmly than ever in the existence of the weapons of mass destruction. And this is probably what also happened to the CIA analysts. Another case was that of those groups of millennialists that believed the world would end on the 1st of January, 2000. Some researchers conducted interviews with some millennialists months after the turn of the century and found that their beliefs in the inevitable and soon to come end of the world were not only still alive, but that they had assumed the fiasco of the 1st of January of the year 2000 as a premonitory sign of the coming of the end of the world, somewhat similar to the CIA analysts that took the lack of evidence of the presence of weapons of mass destruction as a sign that the Iraqi government was hiding them.

Social modulators. Studies on the social dynamics of misleading narratives have shown the importance of social interactions in the unfolding, diffusion and acceptance of such phenomena. To begin with, simple communication between two persons modulates the information that will be borne in mind when evaluating the messages. Various studies

have shown that the information that appears in a conversation between two persons will be that which determines the outcome of the discussion, regardless of the relevant prior information that each interlocutor possesses regarding the subject discussed. All the information that is not mentioned disappears from the equation and is completely ignored. When there is no reason for the hostile attitude, people tend to want to agree with any interlocutor about what they're talking about, and will do anything possible to come to this agreement, even negating their own knowledge.

In the evaluation of an event, the social context in which we experienced it is also important. The opinion of the people with whom we experienced an event can modulate our evaluation of the event. In one study, the researchers showed a documentary to groups of just a few people. Some days later, the participants responded individually to questions about the documentary. A week later they answered the same questions again, but this time after having seen falsified answers from their companions. In spite of having correctly answered the first session of questions, 70

As regards the specific evaluation of misleading narratives, various studies have identified the importance of what are known as "weak ties". Weak ties are those established with acquaintances that share our ideology or group, contrasting with the ties established with friends or families. Weak ties seem to have more importance in the diffusion of fake news, in particular, and of news in general. People seem to exclude tight- knit familiarity in the evaluation of the truth of a news item, and consider that news coming from acquaintances with whom we share a group or an ideology are somewhat more reliable, given that they don't carry the familiarity bias. On the other hand, recent studies have demonstrated that once a version of an event has been established via weak ties, it's very difficult to change that version.

The risk of consolidation of a particular story via social ties, weak or strong, is the consolidation of a partisan view of history or of the contemporary events of a particular community or society. This consolidation phenomenon obviously existed in pre-internet times, but the new digital ecosystem has amplified this trend. Currently, more than 50

5.5 Conclusions

Narratives are a basic component of human cognition by providing meaning to our experiences. Therefore, narrative-based representations should be approached as meaningful representations of events, rather than reliable representations of events. This has important consequences for social-cognition research. In the case of values, for example, the role of narratives as meaningful representations is critical to understand the dynamics of social phenomena, such as radicalism or misinformation sharing, as I have shown in this article. Hence, even if social neuroscientists have been reluctant to use the notion of narratives in their theoretical frameworks and experimental paradigms (Willems, Nastase, and Milivojevic 2020), the fact is that it must take a central role in the future of the discipline.

.

Chapter 6

Narrative Art Interpretation
Luc Steels

Abstract

The paper clarifies the narrative view on art interpretation with a concrete example of a painting by Caravaggio and it explores the implications of this view for building theories and mechanisms for dealing with meaning and understanding in AI systems.

Keywords

Human-centric AI, art interpretation, art production, meaning, understanding, narratives, narrative networks.

6.1 Introduction

The concept of *narrative* is used in many domains of inquiry concerned with humanistic issues(J. Bruner 1991) and it is also central to the theories of meaning and understanding that are being developed and used in semiotics, art history, art criticism and art education. In its most basic form, a narrative consists of a description, in other words a model, of a series of events and their temporal and causal relations. Normally, a narrative model goes far beyond a mere list of events. It also includes many other relations, like spatial or hierarchical relations, actors and entities, and the roles they play in events, properties of events, the general context, the motivations, deliberations and intentions of the actors, a viewpoint, an ideological framing and much more.

Works of art, such as paintings, songs, theatre pieces, films or novels, are examples of *semiotic representations of narratives.*(Eco 1975) A *semiotic representation* consists of signs. A *sign* is a relation between a form (the *signifier* or meaning-carrying element) and

a meaning (the *signification*). For example, an image of a flying lark in a painting can signify a soul ascending to heaven. Signs are often conventional, in the sense that there is an arbitrary relation between form and meaning. For example, the word for 'lark' in Italian is 'allodola' and in Japanese 'hibari'. In art works there is often an iconic relation between form and meaning. For example, the signifier of a lark can be a flute playing a melody similar to that of a lark.[1] A set of signs with a definition of their constituent form-meaning relations is known as a *code* defined in terms of vocabularies and grammars. An alternative equivalent term for sign is *symbol*.

A semiotic representation of a narrative is always incomplete. It assumes that the viewer or listener has the necessary background knowledge not only about the historical and cultural context and events in the narrative model but also about the code that the narrator is using.

6.1.1 Narrative interpretation

During the interpretation of an art work, a viewer (resp. listener or reader) constructs or activates in turn a rich mental model based on perceiving, decoding and interpreting the signifiers of the art work, augmented with experiences from personal memory and semantic background knowledge. In theories of art, this rich multi-faceted mental model is called a *narrative interpretation*. The descriptions that constitute a narrative interpretation can be schematized as a graph with nodes for the various entities, concepts, and the relations between them. I will call such a schematization a *narrative network*.

Viewers construct narrative interpretations in order to make sense of the various inputs that they get, starting with the sensory experience of the art work itself, augmented with suggestions made by the title, information from a catalogue, the context of the exhibition, previous work by the artist, knowledge of the historical figures or events being depicted, knowledge of the artistic movement the artist is a part of, etc.

The influential art historian and semiotician Erwin Panofsky has introduced six different perspectives of description as the main ingredients of narrative interpretations of visual art works(Panofsky 1939,1972). Similar suggestions and analyses have been made for music(Copland 1939) (Minsky 1981), film(Monaco 2000), literature(M. Bal and Boheemen 1997) and other artistic media, although in the remainder of this paper I will mostly focus on visual art works. The six Panofskyan perspectives are:

1. *Sensual*: These are the descriptions that are directly observable or derivable by visual and syntactic processing: the surface, the type of paint, the different colors, the color segments, lines, textures, contrasts and volumes, their aggregation into larger structures and more complex visual entities, called motifs.
2. *Contextual*: These are descriptions of the context: the time the work was made or shown, the exhibition and its theme, the building, the title, other works shown in

[1] As is the case in the tone poem 'Im Abendrot' by Richard Strauss which is about the end of life.(L. Steels 2021)

the same space, the artist.

3. *Factual*: These are descriptions of the real or imaginary events which the art work is about, what actors and objects are involved, which roles they play, and the properties and causal and temporal relations between the events.
4. *Cultural*: These are descriptions of the historical, political or religious significance of the situation depicted in the painting.
5. *Expressive*: These are descriptions of the moods and affects suggested by the gestures, actions and facial expressions of the different characters and of the scene as a whole, amplified by the choice of colors, shapes and perspectives.
6. *Intentional*: These are descriptions of how the events are framed, highlighting some aspects as more important than others, identifying the intentions of the artist and the moral and ideological stance implicitly assumed.

These different perspectives are not experienced one after the other. They are interlaced and emerge in parallel, one influencing the other.

Different artistic media not only use different material forms for signs, they also invite the viewer to explore different perspectives in more or less depth. For example, abstract visual art and (abstract) music, such as Bach's Wohltemperierte Klavier, focus on the sensual and contextual perspective (perspective 1-2) without any expectation that the viewer goes much beyond it. The joy of perceiving the art work then comes from pleasurable material sensations, for example the experience of a color or a sound you find beautiful, and from recognizing motifs and how they have been transformed, expanded and composed. In the case of figurative art or figurative music, such as Vivaldi's Quattro Stagioni, viewers search primarily for meaning (perspective 3-6), seeing beyond or ignoring the visual/auditory appearance, just like when you look through a window and are no longer distracted by the window frame or glass pane but focus on the world behind it. Narrative texts (novels, theatre) also emphasize the meaning perspective and the reader/listener is invited to reconstruct the events being evoked, the characters and their properties and the motivations and intentions of the characters as well as those of the artist.

Literature studies further make a useful distinction between *story* (also called plot), *fabula*, and *narrative text* or *narration*.(M. Bal and Boheemen 1997) The fabula is a description of the actual facts (comparable to the factual and cultural perspective)[2] whereas the story is the transformation and shaping of the fabula to include the cultural, expressive and intentional perspectives. For example, an author typically transforms a historical figure into a character, a 'personnage', with some of the personality traits amplified. In addition, the author highlights certain events, dramatizes them, or perhaps even transposes the historical context to another era. The narrative text is the semiotic representation of the story and its underlying fabula. It is the actual text with specific choices of words, grammatical constructions and text structure.

[2]This term is somewhat confusing because in many European languages (e.g. Dutch, French) a 'fabel' or 'fable' is an account of fictive events.

Similar distinctions are also being made for visual art works. For example, in the case of a portrait, the fabula is the image of a person's face in normal attire and in a setting where this person is usually found, whereas the story is how this person and what s/he has done is interpreted by the painter. To express this story, the painter may deform the facial characteristics of the person entirely, zoom in on certain parts of the face only, use non-natural colors, poses, dress and context, all designed to bring out the motivations, emotions or historical setting of the depicted person.

6.1.2 Narrative design

In the process of making an art work, artists construct narrative networks as well.(Labov 2006) Now the network functions as a *scaffold during the design and fabrication process*. To distinguish these networks from narrative interpretations I call them *narrative designs*. Structurally, they are entirely similarly to narrative interpretations, i.e. they are rich multi-faceted descriptions from different perspectives. But they are now based on the artist's imagination, personal observations of reality, historical facts, and reminiscences of other art works, all coloured with the artist's own perspective, opinions, beliefs and moral stance.

The concrete art work, such as a painting hanging in a church or a narrative text published as a novel, expresses selected elements of this narrative design using the means available in the chosen artistic medium. The artist introduces signifiers (meaning-carrying forms) that trigger the process of narrative interpretation. For example, painters might depict characters performing a particular action using brightness contrasts to highlight what they see as important and choosing colors, gestures and facial expressions that convey an emotional and moral stance towards the depicted action. An artist is in this sense acting like a cognitive engineer,(Donald 2006) trying to influence the interpretation processes of the viewer, for example, by introducing a focal point that influences the viewer's initial eye gaze and subsequent shifts in attention.

Viewers try to detect the signifiers the artist introduced and to guess the deeper interpretations. They need to engage not only in bottom-up processing of sensory experiences using pattern recognition and structural analysis and use their knowledge of the codes used by the artist to map patterns and structural features to possible meanings. They also need to actively project on the art work their own past experiences and their cultural and historical background knowledge. There is no guarantee that a viewer's narrative interpretation is equal to the artist's narrative design, in fact that will never be the case. Viewers autonomously identify their own signifiers and project interpretive descriptions which could be entirely different from those intended by the artist.(Eco 1979) Some artists (particularly musical composers) even claim that they do not want to convey meaning. The meaning has to be supplied entirely by the viewer.[3]

[3]For example, Stravinsky claimed: "I consider that music is, by its very nature, essentially powerless to express anything at all, whether a feeling, an attitude of mind, a psychological mood, a phenomenon

6.1.3 Dimensions of Understanding

The interpreter of an art work strives for a narrative interpretation to maximize various *dimensions of understanding* so that it becomes:

(i) *coherent*, by resolving as much as possible ambiguities and inconsistencies,
(ii) *grounded*, by resolving references to objects and events in reality or in the fictional world of the narrative,
(iii) *integrated*, by trying to fit the different interpretational fragments together into a single coherent whole,
(iv) *compatible*, with prior experiences and knowledge in personal memory, and
(v) *relevant* to questions the interpreter had before or during the interpretation process.

The more this is the case, the more a narrative interpretation is felt to be satisfactory and the interpreter will say that s/he understands the art work. Understanding is therefore a stage in narrative interpretation in which the five dimensions listed above obtain sufficiently high values. It is in principle possible to quantify these dimensions and define an index of understanding. But we cannot expect this index to be universal nor totally objectively definable. Moreover, new descriptions may always be added to deepen understanding and conversely new input may lead to a shift in interpretation, causing puzzlement rather than more understanding. For example, a changing cultural 'zeitgeist' or personal development of the viewer may throw new light on a work, or a fictional person may be recognized to be that of a real person bringing in a flood of new associations and perspectives. The creator has to supply enough detail with enough clarity and shared code to make the interpretation doable although challenging. At the same time s/he avoids that the viewer gets distracted or bored with superfluous details.

6.2 An example of a narrative interpretation

To illustrate better the narrative viewpoint on art, I will now discuss a concrete example. As a starting point, I invite the reader to first take a close look at a painting by the famous Italian baroque painter Caravaggio made in 1602 and shown in Figure 6.1. Looking at this Caravaggio painting, what do you see? Here is one account, but another person might follow a quite different path to understanding, for example, starting with the title. But suppose you do not know the title or the painter or anything about what might be depicted on the painting. You would however see colors, edges, segments, lines, brightness contrasts, but quickly you will zoom in on what is depicted (the factual meanings). You recognize figures against a dark background, suggesting that the scene takes place during the night. Most probably your eye gaze is first drawn to the two figures a bit left from the middle. The right figure has just embraced and kissed the left figure. The left figure is leaning backwards, seemingly trying to avoid the embrace. Further left we see a man with

of nature, etc." See (Stravinsky 1935), p. 53.

Figure 6.1: Painting dated 1602 by Caravaggio (1571-1610). National Gallery of Ireland.

his mouth wide open, stretching out his arms and hands. On the right of the embracing figures we see two other characters dressed in some type of black metal harnass. Glancing further to the right, we see another figure dressed in regular clothes and to the left of him we see, hardly visible, a third figure with a metal helmet.

These observations are already enough for most Europeans to hypothesize that the harnassed characters must be soldiers who come to arrest somebody. And if you are familiar with catholic religion, you probably already realized that this scene depicts a famous episode from the passion story, namely the arrest of Jesus Christ, which will be followed by his condemnation, crucifixion, death and resurrection. The passion story is the most central narrative of catholicism and was known intimately to everybody in the society in which Caravaggio lived. It has been the subject of yearly rituals, numerous paintings, musical evocations and sacred texts. Those who viewed this painting in the early 17th century, possibly in a church or chapel, would recognize the scene instantly and know many more details that are not depicted here. They would have believed unquestionably

that the arrest had really happened and felt strong empathy towards the figure of Christ who was going to go through a terrible tortuous process eventually leading to his death.

The hypothesis that this painting shows a scene of the passion story is born out by the title of the painting: 'Presa di Christo nell'orto' - the arrest of Christ in the garden. Once we know that title, several additional details fall into place. The garden must be the garden of Getshemane where the gospels locate the arrest. The man who has given the kiss must be Judas, the apostle who betrayed Jesus for money, and the one who is kissed is most definitely Jesus. Judas wears a kippah, a cap made of cloth, emphasizing his Jewish allegiance. The depiction of Jesus is further confirmed by the fact that his face resembles a well known icononic image of Jesus on a relic veil supposedly taken by a woman named Veronica during a later stage in the passion events.

The figure on the left of Jesus must be another one of the apostles. Contemporaries of Caravaggio would have easily recognized him as being Johannes, because he is described as being present in the garden with Petrus and Iacobus. He was much younger than the other apostles and therefore usually depicted without a beard. According to the gospel, Johannes fled the scene after being stripped from his cloak. He was sounding the alarm about the arrest and escapes here in order to tell what has happened. The soldier in the middle of the painting is grabbing Christ's throat. The other one is grabbing the cloak of Johannes with his two hands. The three soldiers and the bystander are packed tightly together, pressing against Jesus and Judas. This gives the impression of a mob coming to take Jesus prisoner. Notice how quickly and smoothly we went from recognizing a few figures and their poses and gestures to an interpretation in terms of well known figures, thus moving from factual to cultural meanings.

The more we look at this painting the more we see. Let us focus for example on the expressive meanings. Every figure has a facial expression and posture characteristic for a particular mood or a particular role in the action. The hands by themselves tell part of the story. Johannes is clearly in panic and crying for help with his hands stretched out. Christ looks distraught and puzzled, seemingly asking Judas why are you doing this? He is wringing his hands, a symbol of distress in the face of dilemma. Judas seems to realize too late that he did something terrible and is perhaps showing remorse. The eyes of the soldiers are not visible, as if they need to remain anonymous. The bystander on the far right looks curious but not in the same state of panic as Johannes.

There are many other remarkable details that become apparent on further examination, showing the incredible mastery of Caravaggio as a painter. The positions of the different heads form a V-shape, creating a focal point with the two lines crossing right at the region where Judas has kissed Jesus. This is why our first gaze is drawn to that point. It is the central topic of the painting. Since the background is completely dark the figures stand out sharply and they prominently reflect a light source coming from somewhere behind the viewer. Patches of light fall on the faces of the main protagonists (Jesus, Judas, Johannes) and on their hands, so that our gaze is naturally drawn to them. The draping of the clothes, the details of the central soldier's armour, the detailed representation of

the hands and faces are all painted with an incredible skill in strong vivid colors.

But who is the rightmost figure dressed in ordinary clothes? He is holding a lantern throwing light on the scene and seems an anonymous bystander. It has been argued that he could be another one of the apostles, namely Petrus. But Petrus (still according to the gospels) took defensive action by cutting off the ear of a servant, an action which is not shown at all in this painting. So it is probably not Petrus. Remarkably, it turns out that the anonymous curious bystander on the right depicts Caravaggio himself. We know this because there are known portraits of Caravaggio and because the 'Taking of Christ in the Garden' is not the only painting where Caravaggio introduces himself. For example, his famous painting of David and Goliath now at the Galleria Borghese in Rome, shows the biblical theme of the young David who has defeated and cut off the head of the giant Goliath. The head, which exhibits an expression of anguish and pain, also turns out to be that of Caravaggio. Caravaggio had become obsessed with beheadings after a brawl on the streets in Rome during which he had killed a man. He had been condemned to be beheaded - forcing him to flee and worrying the rest of his life that he would once be captured and beheaded. It is not far fetched to interpret this painting of David and Goliath as a recurring bad dream of Caravaggio and a plea to the powerful cardinals who were his patrons to pardon his murder and save his life. The lantern is another signifier that adds weight to the hypothesis that this person is Caravaggio himself. The lantern was a common symbol for painters at the time, the emblem of the Roman painter's guild. It signifies that a painter tries to see what is in the dark or no longer observable and can make the invisible visible. But the lantern was also a symbol for betrayal in Roman iconography.

By putting himself in his own painting as an observer and reporter of the scene, Caravaggio creates a meta-level. He asserts for himself a role similar to that of the apostle Johannes shown on the symmetrically opposite left side, namely to propagate the true faith. Paintings were seen around the turn of the 17th century as a powerful tool to combat the protestant reformation which was making heavy inroads into traditionally (Roman) catholic areas. Protestant preachers like Luther or Calvin had pretty much banned the use of images, emphasizing words and a literal interpretation of the bible. One of the key ideas of the counter-reformation was to encourage the creation and use of images that believers could contemplate during their prayers or devotions, helping them to imagine vividly the sufferings of Christ, the resurrection, the after-life in heaven and hell, and other key subjects from the scriptures.

Another intriguing element is the arm of the front soldier in black metal with a band reflecting white light. This element is right in the center of the painting and strongly draws our attention. What is its meaning? The metal armor symbolizes a mirror held up to the viewer. This suggestion is not as far fetched as it seems because Cavarragio used a similar depiction of a mirror as a black surface reflecting the dark black background, with a light patch, in another painting, namely 'Martha and Maria (Magdalena)'. In that painting the mirror is symbolic for vanity. But here, it projects the viewer into the picture. The arm (and hence the mirror) is placed before the body of Judas and when

the viewer looks at the painting his own image fuses with that of Judas, the ultimate sinner.(Apesos 2010)

It is difficult for us today to fully understand and experience this painting the way people at the time of Caravaggio did. Most people are no longer familiar enough with the details of the passion story and if they are, they do not believe it to the extent that was common in 17th century Europe. But we can still experience the tremendous force of the art work which calls upon so many different levels of meaning. These meanings are triggered partly by visual elements but also by memories and cultural knowledge. Some of these meanings are symbolic and make the viewer think and reflect on his faith whereas others affect psychological states with feelings of empathy and fear for what is going to happen. They are hard to express in words. The title of an art work, its exhibition context and what is written about it by curators help the interpretation process. Indeed, the role of art historians and curators is to fill in the context so that we can still have a rich experience of an art work, even if it was made centuries ago.

Other painters (or Caravaggio in other paintings of the same scene) have highlighted different actions and tell the story in different ways. For example, a representation by Albrecht Duerer of the same scene highlights the kiss by Judas as well as the action of Petrus whereas a representation by Arpino puts the kiss in the background and highlights both the action of Petrus and the fleeing of Johannes (see Figure 6.2). The JESUS-BEING-TAKEN-PRISONER-FRAME would be known by all viewers at the time of Caravaggio and also the side actions would be recognized instantly.

6.3 The hermeneutic spiral

A recurrent theme in studies on interpretation is a paradox known as the *hermeneutic circle*: *To understand the whole we need to understand the parts but to understand the parts we need to understand the whole.*(Gadamer 1975) Only for the most simple situations is there an instant recognition. Instead we usually experience a gradual process, flipping back and forth between trying to grasp the total picture and identifying and interpreting individual signifiers.

There are several reasons why the interpretation process is gradual:

1. The sensory perception of art works takes place *sequentially*: You cannot see a visual art work in one glance but you have to scan it. You have to read a text one word after another. You can only listen to music as it unfolds in time.
2. Semiotic representations have inherently a lot of *ambiguity*, in the sense that the same set of features can trigger different signifiers and the same signifier can have multiple interpretations. For example, a line segment can be part of different objects, or a particular facial expression may be a sign of laughter but also of crying. The resolution of ambiguities can often only take place when additional elements have been processed and the overall setting has become clearer.

Figure 6.2: Alternative representations of the Arrest of Jesus. *Left:* Version by Albrecht Duerer from 1511. This very influential version clearly inspired Caravaggio's, illustrating how certain depictions of an event became iconic and were transmitted for decennia from one generation to the next. The kiss of Judas is shown centrally but also Petrus cutting off the ear of a guard, which is not shown in the Caravaggio painting. The fleeing of Johannes while losing his cloth (described in the Marcus version of the passion story) is not shown by Duerer. *Right:* Version by Cavalier d'Arpino from 1597. Judas is hardly visible, Petrus cutting off the ear is shown, as well as Johannes (in the left corner) losing his clothes while fleeing. The soldier to the right of Jesus is depicting in a similar way. Notice also the lantern in the two paintings.

3. Semiotic representations have a lot of *indeterminacy*: They highlight certain aspects of the story but provide less detail for others, either because of the way the author wants to frame the story, or because many details are irrelevant or assumed to be known as general knowledge or from information obtained later, or because the author creates suspense by leaving out details to stimulate the imagination. For example, we may already know that a painting depicts a face but not yet to which person this face belongs. Maybe we will never know or we do not need to know because the painter simply wanted to paint 'a face'.

The gradual process of deepening understanding is called the *hermeneutic spiral*: Starting from an initial examination of some elements (with a lot of ambiguity, uncertainty and indeterminacy) a human interpreter constructs the first hypotheses of the whole, which then provide top-down expectations to be tested by a more detailed examination of the

same or additional elements, leading to a clearer view of the whole, which then leads back to the examination of additional parts, etc., until the interpreter has reached a satisfactory level of understanding.

6.4 Steps towards computational models

Everything mentioned so far is well known from the literature in semiotics and art and widely accepted. We now turn to the question in how far the highly complex mental processes needed for art interpretation and art design can be modeled computationally, as a path towards developing mechanisms that are relevant for giving some notion of understanding to AI systems in general. Although there has been very little work on this so far, there are many components and the big challenge is to integrate them.

Knowledge Sources

Clearly, interpreting art works (not to mention the design and fabrication of art works) has to rely on a wide scala of *knowledge sources*. Fortunately, massive recent work in AI has seen steady advances for many of them. Knowledge sources can be classified in two ways.

The first classification is based on what level in the mappings from observable forms to meanings a knowledge source contributes to. Three types are relevant for art interpretation:

- *Knowledge sources for sensory processing* treat the raw input data in order to derive entities, features of entities and hierarchical structure. In the case of visual sensory processing, low level algorithms provide approximate information about color segments, edges, textures, shapes, light sources, etc. Pattern recognition algorithms attempt to recognize objects, components of objects, properties of objects, how they move, etc. There are by now thousands of knowledge sources for sensory processing and they are growing rapidly. Many of these sources are made available using open platforms such as OpenCV (Open Source Computer Vision Library - `https://opencv.org/` and maintained and expanded by a world-wide community.
- *Knowledge sources for decoding* use models of a code in order to find out which possible meanings relate to features and patterns. Semiotic representations use codes which can get extraordinarily sophisticated. The most obvious example is language where the code is defined in terms of phonologies, morphologies, lexicons, grammars, semantics and pragmatics. Particularly for human natural language processing (NLP) there have been rapid advances in the state of the art made available through open platforms such as Spacy (Industrial-Strength Natural Language Processing - https://spacy.io/).
- *Knowledge sources for semantic processing* are concerned with expanding narrative networks by making inferences based on knowledge bases. Also here there have been massive advances the past two decades, particularly in relation to the emergence of

the semantic web.(Antoniou and Harmelen 2008) We now have large encyclopedic knowledge repositories in the form of knowledge graphs in order to answer questions relevant for reconstructing the fabula of an art work, filling in the cultural and historical context, collecting information on the painter and the artistic movement of which s/he is a part, and much more. (Boer et al. 2013) Access to this 'intertextual encyclopedic knowledge', that Umberto Eco considers a prerequisite for understanding of narrative texts,(Eco 1979), p. 208 is no longer completely beyond the reach of AI systems.

The second classification is based on how a knowledge source achieves its purpose. There are essentially two approaches: one based on machine-learning and the other based on design, knowledge acquisition from human experts and abductive learning:

- The machine learning approach produces statistical models based on training with a (large) data set. The dataset is either human-annotated, enabling supervised learning, or bare, requiring statistical induction, for example based on successful prediction or completion of a pattern.
- The design approach is based on human-designed functional models. This approach relies on the availability of human expertise which is acquired by an analyst (called a knowledge engineer) from a human expert in an interactive knowledge acquisition process or on learning mechanisms (such as explanation-based learning) that autonomously acquire functional models through abduction and feedback from testing out a model in concrete circumstances.

There are of course also mixed forms, known as *hybrid approaches*, that combine the two. For example, some AI systems learn heuristics using neural network techniques for coping with the combinatorial explosions generated by a human-designed model. This approach is the basis for alphaGo or for recent work in the acquisition of heuristics to scale up construction grammars.(Van Eecke, Nevens, and Beuls 2022)

An example of a knowledge source for sensory processing is to detect the focal point of a painting to which the human eye gaze would normally be drawn when seeing a painting for the first time.(L. Steels and Wahle 2019) Dozens of methods are known for this task, some of them, such as MSI-net, rely on neural networks (contextual encoder-decoder networks) trained with annotated data.(Kroner et al. 2019) Others, such as the Montabone and Soto Visual Saliency algorithm(Montabone and Soto 2010) rely on the detection of color constancy, pair callibration, depth continuity segmentation, and many other mathematically derived image features.

There is already a very large set of knowledge sources for language processing based on statistical induction of models of language codes from language data. This is for example the case for most of the tools contained in the spaCy platform. But there are also expert-designed grammars based on linguistic analysis. These grammars often start from an expert-designed kernel that is automatically expanded based on specialized learning strategies.(Beuls 2017)

Data-oriented techniques, particularly distributed semantics and statistical models are currently being explored extensively for building systems that can function as knowledge bases. Recent examples are BERT and GPT3. Alternatively, many knowledge graphs have been built or extracted from texts that are carefully curated from contributions by human experts (as is the case for Wikipedia).

It is important to be aware that there are no perfect knowledge sources. Instead, a knowledge source typically works well for a specific subdomain and even then it will provide outcomes with only some degree of certainty. There are several reasons for this: If the knowledge source relies on statistical models and training, its competence will depend on the statistical distribution of the data and the machine learning methods that have been used. For expert-curated knowledge sources, we often see differences in opinion between experts and incompleteness due to the vastness of human knowledge and the unavoidable limits on what an individual can know.

Compositional use of knowledge sources AI research has explored two fundamental models for combining knowledge sources. The first one is based on the notion of a *pipeline*. Different knowledge sources are chained with the output of one source being the input to the next one and so forth. The alternative is a *blackboard architecture* in which the different knowledge sources act like experts that read information on the blackboard and then write more information based on their own specific competence. Knowledge sources become active when they see relevant patterns on the blackboard (this is called pattern-directed invocation).

The blackboard model, pioneered in the late 1970s for speech understanding(Erman et al. 1980) and knowledge-based expert systems, appears better suited to implement the hermeneutic spiral because it is difficult (if not impossible) to define in advance a rigid sequence in which knowledge sources should be applied. Moreover during the understanding process information needs to flow both in a bottom-up and top-down manner across levels of description. In a pipeline model uncertainty produced by one knowledge source propagates and gets amplified to the next in the pipeline, whereas it is much more effective to weigh the evidence of different knowledge sources to achieve greater robustness and reliability.

To make the blackboard architecture a reality we need the following:

1. We need a datastructure that records the progressively deepening set of descriptions which make up a narrative interpretation. I will call this datastructure a *transient narrative network* because it is mathematically speaking a network, in other words a graph, similar to the kind of semantic networks already commonly used in AI., and because the network will be dynamically changing as new concepts and nodes are added due to additional input or the invocation of additional knowledge sources.
2. We need a way to orchestrate the many processes that become active as the hermeneutic spiral takes its course. Here I to consider a *task-based control architecture* commonly used for organising heuristic search processes in many areas of AI and robotics

since the 1960s.(Forbus and Kleer 1993)

3. We need a mechanism that can help to inject top-down expectations and regulates which knowledge sources become active in order to confirm the expectations, check whether default assumptions hold, or fill in more details. Here I suggest to consider the notion of a *frame* and of *frame systems* that emerged in AI in the mid-1970s.(Minsky 1975)
4. We need an attention mechanism that determines which part of the narrative network should be explored in more depth or which additional sensory observations might be useful, for example, which area of the painting should be examined in more detail. Here I suggest that we should explore mechanisms based on the notion of self-organizing systems pioneered in AI in the 1980s, specifically a *self-organizing attention mechanism* that is maximizing the dimensions of understanding mentioned earlier.(L. Steels 1991)
5. This attention mechanism will have to rely on a series of internal measurements of the understanding process: (i) how far ambiguity, incompleteness, underspecification and incompleteness is being reduced, (ii) how information has become integrated, meaning in how far network fragments could be connected to each other and how connections between different levels of the narrative (the fabula, story, and narration) could be established, and (iii) how far the distance towards a satisfactory narrative closure could be bridged. Narrative closure happens when the key questions for which the narrative is being developed have been answered.

6.5 Conclusions

What conclusions can we draw from this foray into the domain of art interpretation? Clearly, to understand what art is about we have to look at the effect it seeks to have on the viewer. An art work triggers a process of perception and meaning creation progressively leading to deeper and deeper understanding. Which meanings a particular viewer evokes depends on many factors, including his or her knowledge and past experience, familiarity with the codes used by the author, and the time and effort s/he is willing to invest.

It is futile to look for an objective ground truth, a canonical way of interpretation that everybody is supposed to reach. Some people walk into an exhibition, spend a few seconds on a painting and go on to the next. They will see very little and soon forget what they saw. Others will spend minutes, come back later, read about the painting, the painter and the period it was made, try to figure out the broader context and the cultural ecosystem in which the art work first thrived. They will revisit the work time and time again until the painting becomes a trusted acquaintance that gives pleasure, insight and guidance. It is simply astounding that an art work like Caravaggio's 'Presa del Christo nell'Orto' still fascinates us four hundreds years after it has been painted and continues to be the subject of books, documentaries and exhibitions. There is an almost magical extraordinary power in great art works, precisely because the experiences and meanings they evoke are endlessly

rich and every epoch viewers re-interpret and give new or different meanings to a work.

A second conclusion is that it is clearly completely beyond the current state of the art in AI to achieve even a small inkling of what the human experience and understanding of art or the construction of new art works is like - despite some claims to the contrary. We need significant advances in the many knowledge sources that have already been developed within AI. We also need advances in the general architecture of AI systems that understand.

The statistical machine learning approach that is currently *en vogue* is going to help in this endeavour but it is not the only route we should follow for two reasons: (1) The machine learning approach targets prediction using observable data as sole input. But prediction is not the same as understanding. For example, it is surprisingly possible to predict to a remarkable degree the next word in a sentence, given enough statistical data, even *without* knowing what the sentence is about or without mastering a causal model of how language expresses meaning.(Devlin et al. 2019) But the purpose of language is communication and listeners primarily try to understand a sentence rather than predict how it continues. This is true for art works as well, particularly for figurative art works that are making a point. Moreover, (2) most of the information needed to reach understanding of an art work is not in the work itself but must be projected by the viewer based on semantic knowledge and prior experiences. This implies that a supervised training scenario can only be used if we can annotate art works with the rich set of descriptions that narrative interpretation or design require and generate.

Finally, these conclusions do not mean that AI is forever irrelevant to art, on the contrary. As I very briefly described in the paper, there are already many useful tools to aid in art interpretation and there have been very significant advances in the many functionalities that play a role in understanding (in computer vision, pattern recognition, language processing, knowledge representation). In turn, the study of art interpretation and narrative design gives us ideas on how to push the state of the art in AI so as to incorporate better the notion of narratives and it is a path towards tackling the fundamental issue of meaning and understanding in human-centric AI. Specifically in addition to the already existing knowledge sources (which could still be improved) we need to explore others that are relevant for the interpretation of narratives, such as semantic frame extractors, pragmatic analyzers, etc.

Acknowledgement The author thanks Oscar Vilarroya (IMIM and UAB Barcelona) for extensive discussions over the past years about meaning and understanding in various art forms.

.

Chapter 7

Pragmatics of Narration with Language
Anna Morbiato

Abstract

This paper explores how research conducted into pragmatics and cognition in the past fifty years can deepen our knowledge on natural languages, narratives, and ultimately language processing. Studies along this line were aimed at exploring how humans organise linguistic forms and expressions under the influence of several types of non-linguistic factors - that had for the most part been neglected by core linguistics - including the speaker's communicative aims, perceptual experiences, stance/viewpoint, as well as cultural and social aspects of communication. One of the first and most effective efforts towards this goal is the Pear stories Project launched by William Chafe, which revealed crucial insights into the relation between languages, cultures, and how humans perceive, experience, and retell the same story, serving as foundation for much progress in understanding both spoken and written language. The paper ultimately explores how some of the fundamental pragmatic elements that characterise narratives are present and in fact heavily influence the ways languages structure basic linguistic elements, such as clauses, sentences, and texts.

Keywords

Narratives, narrations, linguistic pragmatics, information flow, salience, perspective taking, speech acts

7.1 Introduction

This paper explores how research conducted into pragmatics and cognition in the past fifty years can deepen our knowledge on narratives, narrations in language, and language processing. It offers an overview of but some of the major topics that characterise the realm of pragmatics, including notions such as attention, information flow, contextualisation, inference, world knowledge, topic, focus, coreference resolution, anaphora, viewpoint, salience, perspective taking, and speech acts. Specifically, it aims at illustrating how these and other fundamental features that characterize narratives are present and in fact heavily influenced the ways in which language(s) structure basic linguistic elements, such as clauses, sentences, and, ultimately, texts. Finally, it opens some perspectives on how pragmatic aspects of narratives may impact social dynamics from a sociolinguistic perspective.

Following the terminology of narratology (M. Bal and Boheemen 1997) we make a distinction between the fabula or factual aspects of a narrative, the story or framing of a narrative, and the narration which is a semiotic representation of the story. Here we focus in particular on narrations that use language, i.e. that rely on a linguistic code.

The core of narratives are one or more events connected with temporal and causal relations(L. Steels 2022). Among the most basic elements of an event are its *participants and their roles*. This is generally captured by so-called semantic roles (agent/actor, patient/undergoer, beneficiary, goal, etc.). Different languages encode such event participants in different ways, some through syntactic roles (subject, object), some through different coding (cases, cross-reference, word order etc.).

Other essential elements of events include the *temporal and spatial settings* in which they occur, as well as the *causal relationships* between them. This type of information may be encoded in a number of means, including tense/aspect system/consecutio temporum ('I went there after he had moved elsewhere'), adverbials ('In 54 BC, Caesar invades Britain'), connectives ('Because of your actions, I cannot leave town'). Moreover, in a series of events, the participants that are involved and the spatio-temporal settings in which the events happen typically change as the narration unfolds: for the communication to be effective, all these elements need to be easily interpreted and - hence - identifiable by the interlocutor. This is well expressed by Foley and Van Valin (Foley and Valin 1984), p.1.:

> When talking about sequences of situations in which the same participants are involved, it is necessary to refer to them in each clause in such a way that they can be identified as being the same as or different from the participants referred to in previous clauses. Moreover, speakers need to signal the temporal relations between situations. (Foley and Valin 1984), p.1.

This quotation introduces some of the main core issues in the study of pragmatics: the *choice of anaphoric means* (i.e., ways to refer back to previously mentioned event participants or settings), *coreference resolution*, *topic continuity* (discourse topics) and shifts,

which will be briefly presented in the next sections. Other aspects of narration that are essential to the adequate interpretation of a narration include the *general context, the speaker's stance, viewpoint, and communication purposes*, the *contextualization of the event* within real or fictional contexts, and other aspects that are often subsumed under the umbrella-term 'pragmatics'. Furthermore, one of the most central factors that governs all language structures and communication is the tension between understandability and economy. Essentially, this tension refers to the fact that we would like to get our message across as clearly as possible with as little effort as possible, enshrined in the Gricean maxims of communication.

The ways in which languages manage to convey information related to the vast array of aspects mentioned so far constitute one of the most fascinating - and complex - realms of study for scholars. However, not all theoretical approaches to language agreed that these aspects should be investigated within the discipline of linguistic. For example, the so-called 'generative approach' tended to marginalise, and even cut out, all that was connected to language in use, variation, or communication purposes. But why were these aspects neglected or somehow left aside for some time in mainstream linguistics?

In the second half of the last century, linguistics was deeply involved in trying to find a deep structure common to all languages, therefore eliminating all those 'superficial' (and hence superfluous?) elements that render languages different and mutually unintelligible. Linguists working towards that enterprise were pursuing the search of a genetic, universally identical formalisation of language, which could be built up by performing transformations on just "a small, possibly finite kernel of basic sentences" (Chomsky 1956), p. 124, a task that could be ultimately be performed by a machine. Language and its complexity had to be hence reduced into smaller, simpler, and more manageable tasks.

On the other hand, considering pragmatic aspects implied accounting for a much greater amount of complexity: as Fillmore (Fillmore 1976) observed, for its inherent nature, pragmatics encompasses both the syntactic and the semantic components, and integrates them with aspects of language in use and in context:

> [P]ragmatics [...] unites (i) linguistic form and (ii) the communicative functions that these forms are capable of serving, with (iii) the contexts or settings in which those linguistic forms can have those communicative functions. Diagrammatically:
>
> Syntax [form]
>
> Semantics [form, function]
>
> Pragmatics [Form, function, setting] (Fillmore 1976), p.83

In short, if compared to other domains of linguistics, pragmatics has a much wider scope: it offers "a general functional perspective on (any aspect of) language, i.e., as an approach to language which takes into account the full complexity of its cognitive, social, and cultural (i.e., 'meaningful') functioning in the lives of human beings" (Vandelanotte 2009), p. 19. Under this view, a sound and encompassing account of language looked like a

far more complex enterprise, an enterprise not all linguistic approaches would choose to undertake.

7.2 Chafe and the Pear Stories

One of the first linguists to took on this challenge is undoubtedly Wallace Chafe: he saw language as tightly connected with the mind, as well as with human experiences, understanding, and narrative production. Among his first observations is the fact that people *use language and create narratives depending on what they attend to*, as well as what they assume their listener is aware of (W. Chafe 1974). One of his first and most effective efforts towards a deeper understanding of pragmatics and language is the Pear stories project (W. Chafe 1980), conducted under a grant received from the United States National Institutes of Mental Health to investigate relations between languages, cultures, and how humans perceive, experience, and retell the same story.

The Pear Stories Project is based on a six-minute film, with sound effects but no words, designed to present viewers with a range of differing experiences and mnemonic anchors, which might be verbalised in diverse ways and allow elicitation of a wide variety of typologically interesting linguistic constructions. The movie shows a man harvesting pears, which are stolen by a boy on a bike, who then leaves and interacts with other children; later, the farmer discovers that his pears are missing.

Chafe's core idea is that real language emerges only in connected speech, which is characterised by variations in word order, in the choice of topics and perspective taking, in the use of anaphoric means and cohesive devices, of backgrounded vs. foregrounded information, as well as of what to say and what to leave unmentioned. Throughout the years, the film was shown to people from various countries who spoke different languages and belonged to different cultures, and who were asked to tell what happened in it. It is still used by many researchers around the world as convenient stimulus material for collecting natural discourse.

The wealth of studies based on the film reveal crucial insights into the relation between languages, cultures, and how humans perceive, experience, and retell (i.e., create a narrative) the same story (input), serving as foundation for much progress in understanding both spoken and written language. Studies show interlinguistic variability in the ways speakers refer to event participants throughout the retelling of the story, enlightening our understanding in aspects like referential means, anaphors, and coreference resolution (see 2.7). Studies focusing on cross-linguistic differences in narration and storytelling abound as well (W. Chafe 1980), followed by many other studies).

Other insights were gained on the way the focus of attention changes through time in the retelling of the story with the flow of attention paralleling the flow of information (see 2.1): focuspoints of awareness are limited by the built-in limited information processing capabilities of our brains and such limitations are reflected in the structures of sentences,

that generally consist of a given referent or element (anchor) and some new information on that referent (e.g., what it does or what happens to it, see 2.4).

Among the first things Chafe noticed in the adult narratives of the film is that setting and participants tend to be the first thing to be introduced, according to him: "there are different kinds of information which peripheral attention 'requires,' without which the self, as the user of awareness, is uncomfortable and disoriented"(W. Chafe 1980), p. 41. Here in an example from one of the recorded Pear Stories:

(1) The movie opened up on this nice scene, it was in the country, it was oaks, it was seemed like West Coast. (SETTINGS)
There's a farmer, he looks like a uh ... Chicano American, he is picking pears. (PARTICIPANTS). ((W. Chafe 1980), p.27)

Chafe's ground-breaking work *Discourse, consciousness and time: The flow and displacement of conscious experience in speaking and writing* (W.. Chafe 1994) elaborates on these insights and discusses a number of crucial aspects of narrations with respect to awareness, including how linguistic communication regulates the focus of attention, how identifiability and memory activation influence the choice of anaphoric devices and definiteness markers, as well as how language reflects the speaker's point of view, orientation, evaluation and stance (see section 10). The next sections offer an overview of some of the most salient topics that emerge in the book as well as in subsequent literature.

7.2.1 Consciousness, focus, and orientation

According to Chafe, among the constant properties of consciousness that are relevant to language is the presence of both a focal and a peripheral attention, analogous to focal and peripheral vision: "consciousness is the activation of only a small part of the experiencer's model of the surrounding world, not the model in its totality" (W.. Chafe 1994), p. 29. In this paper I will use the term 'attention' instead of consciousness because this is the function of consciousness that is intended.

A limited attention span is reflected linguistically in brief sketches of language he calls 'intonation units', which he defines as a sequence of words combined under a single, coherent intonation contour, that is usually preceded by a pause. Such intonation units reflect speakers' focus of consciousness and attention at the time of speaking.

Chafe's intuitions are supported by neuro-biological findings: short-term memory holds a small amount of information in mind (Cowan 2008) in an active, readily-available state for a short period of time (typically from 10 to 15 seconds, or sometimes up to a minute) . With respect to language processing, in order to understand a sentence, the beginning of the sentence needs to be held in mind while the rest is read. The fact that attention has a limited, constantly shifting span is in turn evident from the observation that speech is produced in a series of brief, prosodically definable spurts, typically between one and two seconds long (Sandra, Verschueren, and (eds.) 2009), p. 137. Panichello and Buschman

(Panichello and Buschman 2021) also found that attention and working memory share the same neural mechanisms.

This has important connections with cross-linguistic evidence supporting Chafe's 'One New Concept at a Time' Constraint: "Conversational language appears subject to a constraint that limits an intonation unit to the expression of no more than one new idea" (W.. Chafe 1994), p.119. in other words, a particular intonation unit may express only one 'brand new' concept, or activate only one concept from the inactive state; a similar theory had been elaborated by Givón: "there exists a strategy of information processing in language such that the amount of new information per a certain unit of message-transaction is restricted in a fashion-say 'one unit per proposition" (Givón 1984), p.202. Similarly, Du Bois had proposed the 'One New Argument' Constraint (Du Bois 1987), p.829. This is expressed by the fact that a predication is often done on a referent the hearer already knows about (identifiable, known, informationally-old, inferable) as in:

(2)	The subject	was charged with manslaughter
	OLD INFORMATION	NEW INFORMATION (FOCUS)

If this is not the case, many languages have constructions that are 'bipartite', meaning contain two predicates, one introducing the new referent (3-4a), the other (3-4b) predicating upon that referent (which is now activated in the hearer's memory and can now be referred to with a weak anaphoric mean, i.e., a zero-anaphor):

(3) (a)	There's a new suspect$_j$	(b)Ø$_j$	police are looking into.
	NEW REFERENT	ZERO ANAPHOR/OLD INFO	NEW INFORMATION (FOCUS)
(4) (a)	C'est ma voiture$_j$	(b) qui$_j$	est en panne.
	NEW REFERENT	ANAPHOR/OLD INFO	NEW INFORMATION (FOCUS)

The focus of attention does not stand still, but is constantly shifting from one item to another. This pattern of a constantly shifting focus against a peripheral background seems always to be oriented with respect to a point of view that functions in the interests of the speaker and listener.

7.2.2 Activation states and cost

As one focus of attention replaces another, the idea of some referent, event, or state may either remain active or become active. This process underlies what is usually thought of as the distinction between given and new information, or activation cost (W.. Chafe 1994), pp. 71-81). Language works best when *the expression of activation cost is listener-oriented*, in which case a given idea is one that is judged to be already active for the listener, while a new idea is one that is judged to have been previously inactive for the listener. A third category of accessible information, i.e., semiactive state, is necessary, according to Chafe, to characterise an idea that is judged to have been previously semiactive for the listener.

Let us consider example (5) from the crime television program The Mentalist, season 5, episode 18.

(5) (a) The victim's Sharon Warwick$_j$ (b)Ø$_j$23
(c) She$_j$ went off of a balcony$_k$, (d) room 914$_k$.
(e) Rigsby's up there$_k$ with the forensics team.
(f) Now off the bruise on the cheek$_j$ and the ripped vest$_j$, we're thinking it's murder.

The speaker is an agent addressing her boss, who has knowledge of the referent 'Rigsby' who is another agent in her team - hence a known entity, as well as of the referent of 'forensics team', that is coherently introduced by a definite article. On the other hand, the boss has as yet no knowledge of the victim, who is introduced postverbally as new information with a full noun phrase, as well as of other details of the victim (age, death) and crime scene (where it happened).

These three activation costs - given, accessible, and new - provide an example of how drawing attention affects language. Given information is typically verbalised with phonetically attenuated material, as when a given referent is expressed with a zero anaphor, e.g., zero$_j$ in (5b), or a weakly accented pronoun, like 'she' in (5c). Recently evoked or ACTIVATED entities are encoded by less-overt anaphoric expressions (e.g., pronouns in English), whereas newly introduced entities are usually encoded with a primary accent and typically with overt forms (e.g., full noun phrases, such as 'Sharon Warwick' in (5a), cf. (Givón 1984) (Figure 7.1.

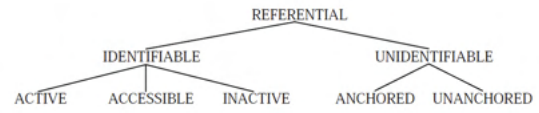


Figure 7.1: Overview of the different types of access of referents from (Pavey 2010), p. 272.

(6) (a) The killer is right behind you! IDENTIFIABLE, ACTIVE
(b) The killer you've been looking for is right behind you! IDENTIFIABLE, ACCESSIBLE
(e) The person we're looking for is a killer you once met. UNIDENTIFIABLE, ANCHORED
(f) The person we're looking for is a young Caucasian. UNIDENTIFIABLE, UNANCHORED

7.2.3 Context and world knowledge

An important source of information which makes it possible for the hearer to easily access/ activate concepts or referents (which are then informationally given) is *context*, as well as *world knowledge*. Example (5) above offers some insights on this (the reader may watch the series scene as well): note that the first NP in (5a) is introduced by a definite article, which in English marks identifiability of the referent That is justified by the fact that the hearer is equipped with world knowledge and contextual cues that suggest that there is likely a victim in the crime scene she is at.

In discussing what is 'contextual', and thus taken as given/presupposed in the hearer's mind, Lambrecht (Lambrecht 1994), pp. 36-37 distinguishes between the *text-externalworld* (which I call context), comprising e.g., (i) speech participants and (ii) speech setting, and the *text-internalworld* (which I refer to as co-text) comprising linguistic expressions and their meanings, in other words, what has been said so far in the previous interaction/text. An example in (5) is the pronoun 'she' in (5c), interpretable thanks to the co-text (the previous sentences, where the referent of 'she' is introduced 'the victim' 'Sharon Warwick'). While text external world components can be taken for granted in the context of the conversation, and thus encoded as given information, the information status of text-internal elements depends on "whether, and how recently, mental representations of these entities have been established in the discourse" (Lambrecht 1994), p.38: If the referent has just been mentioned, it is more accessible; if not, it may need to be reintroduced through a stronger anaphoric means.

7.2.4 The flow of information: topic and focus

Information structure (or 'information packaging' (W. Chafe 1974) deals with the question of how - and specifically, in what order - the speaker chooses to present the informational content of a proposition. There are a number of ways to express the same propositional content, given a verb and its arguments, which may be realised in different positions in the sentence or in different syntactic roles (subject, object, oblique), or enter specific marked constructions, depending on the relative accessibility of entities: highly accessible entities tend to occur early in the utterance, and to be pronominalised; new information tends to occur later on in the sentence, in a flow from given to new.

Consider the following examples:

(7) (a) Trump won that election.
(b) The election, Trump won.
(c) There was an election that Trump won.
(d) This election was won by Trump.
(e) Trump won an election.
(f) It was Trump who won the election.
(g) What Trump won was the election.
(h) (As for) that election, Trump won it.
(i) Trump won it.
(j) He won the election.

Sentences in (7a-j) can all be used to describe the same episode (propositional content), i.e., that a particular person named Trump won an election. Two notions that play a central role in the packaging of information structure are *topic* and *focus*: a sentence topic can be defined as a "matter of [already established] current interest which a statement is about and with respect to which a proposition is to be interpreted as relevant", while focus can be considered as the component that creates "a new state of information in the mind of the addressee." (Lambrecht 1994), p. 118.

Sentences like those above can be analysed in terms of topic and focus:

(8)	(a) Trump	won that election
	(b) The election	Trump won.
	TOPIC (KNOWN)	FOCUS (NEW)

Pavey uses a related set of terms for information structures (Pavey 2010), p, 274.

Presuppostion	Assertion
Topic	Focus
Given/Old	New
Identifiable	Unidentifiable
Background	Foreground

As illustrated in sentence (9)

9	It was Jane who witnessed the robbery.
presupposition:	Someone witnessed the robbery
assertion:	The identity of that 'someone' was 'Jane'.
focus	Jane

7.2.5 Salience

A contextually salient referent with respect to a context/narrative/event is one that the speaker wants to highlight either for linguistic or for extralinguistic reasons. As said earlier, our attentional system has limited capacity and span, hence it needs to filter information depending on its importance/relevance with respect to the speaker's aims. Different languages' grammars have specific devices to encode salience. Consider, for example, an event in which a cowboy is punching a boxer. This event can be described in multiple ways including the following (Tomlin and Myachykov 2015):

(10)	(a)	The cowboy is punching the boxer.	active clause	cowboy is syntactic subject
	(b)	The boxer is/gets punched by the cowboy.	passive clause	the boxer is syntactic subject
	(c)	c) It's the boxer that the cowboy is punching.	cleft sentence	the boxer is pre-posed object of the complement clause

The salience hypothesis assumes that selecting among the structural alternatives is biased by how the event is perceived and conceptualised. For example, the choice of a passive or an active construction in (10a) and (10b) highlights the relative importance/salience of one referent over the other (the cowboy or the boxer).

7.2.6 Discourse topics

As intonation units capture individual foci of attention, each with a strictly limited capacity, one also needs to examine how language is shaped by larger aggregates of semiactive information. Chafe calls these larger coherence chunks 'discourse topics'. "Too comprehensive to be active in consciousness all at once, a semiactive topic must be scanned by a

more restricted focus of active consciousness. When the scanning is verbalised, the result is an episodic or paragraph-like unit of language. The fact that speakers so often begin a topic by providing a setting - an orientation in terms of space, time, participants, and background activity - suggests that information of this kind is required by a well-ordered consciousness" (W.. Chafe 1994), p. 129.

As seen above, the focus of attention does not stand still, but is constantly shifting from one item to another. Let us consider the following paragraph: the main discourse topic is the salient paragraph-initial full-NP referent 'Barack Hussein Obama II', which is referred to in the following sentences through weaker forms (pronouns or zero anaphors) or with shorter noun phrases, like 'Obama' (all anaphoric means are in bold):

11 Barack Hussein Obama II$_j$ (Ø$_j$ born August 4, 1961) is an American politician and attorney **who** served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, **Obama**$_j$ was the first African-American president of the United States. **He**$_j$ previously served as a U.S. senator from Illinois from 2005 to 2008 and an Illinois state senator from 1997 to 2004. (Source: `https://en.wikipedia.org/wiki/Barack_Obama`

7.2.7 Coreference interpretation and anaphoric means

As we said earlier on, when talking about referents, speakers choose how to describe them based on a number of factors, e.g., how well our hearer knows the referent, or whether it was previously mentioned in the discourse or present in the context/situation.

(12) (a) Obama likes knitting.
(b) The former US president likes knitting.
(c) He likes knitting.
(d) That man over there likes knitting.
(e) A man I know likes knitting.

All of the underlined noun phrases could refer to the same person. Anaphora can be defined as a relation between two linguistic elements, in which the interpretation of one (called an anaphor) is in some way determined by the interpretation of the other (called an antecedent):

(13) John$_j$ was sitting at the table. He$_j$ was daydreaming about the weekend
ANTECEDENT REFERENTIAL DEVICE/EXPRESSION

The choice of the anaphoric means depend on the factors given above. For example, a proper name, e.g., Obama in (12) or John in (13), is used if the speaker is sure the hearer has heard of the referent, even if it had not recently been mentioned. The use of a definite noun phrase also implies the speaker thinks the hearer will associate the NP, e.g., the former US president in (12), with the same referent as she has in mind, because that referent is identifiable to them. The choice of anaphoric means is crucial in longer texts: at least every third word in natural discourse depends on referential choice.

7.2.8 Foreground and background information and its coding

The concept of grounding - foregrounding and backgrounding - can be understood by imagining a play, consisting of the main characters and the major development of the story, as well as with the background scenes and elements. The main characters and the main line of the story are "foregrounded", while the supporting scenes and characters are "backgrounded." When a story is told, a distinction is made between the part of text that are foregrounded - important events in a narrative, central points of an exposition, and main characters or entities involved. Backgrounded text, on the other hand, codes events that are less important, secondary items or circumstances, points of reference in time and space.

From a linguistic point of view, foregrounded clauses in narrative discourse tend to narrate past, completed, dynamic events that actually occurred and that can be seen as the backbone of the story: such events are essential to the thematic development of discourse. Backgrounded clauses tend to display durative or stative aspectual traits - they are descriptions and elaborations, and are less important to the story development.

While cross-linguistic studies of grounding have revealed that the semantic features of grounding tend to be universal, the background - foreground distinction is made by grammatical devices that vary from language to language. In general, tense - aspect marking and subordination are major grounding devices. In English, for example, foregrounded clauses tend to use dynamic verbs in finite modes, e.g., in the past tense. Backgrounded text uses non-finite and imperfective verbs, as well as subordinate clauses.

These devices are illustrated by the following excerpt from (Li 2018) (foregrounded clauses are in italics - the text is divided into clauses and marked with lower-case letters in parentheses). The original text is from *The Notebook* by Nicholas Sparks.

(14) A little before noon, (a) *Noah and Allie went in* (b) to prepare lunch. (c) Both of them were starving again (d) because they hadn't eaten much the day before. (e) Using what he had on hand, (f) they *fried some chicken* and (g) *baked another batch of biscuits*, and (h) the *two of them ate* on the porch, (i) serenaded by a mockingbird.

Foregrounded clauses in (a), (f), (g), and (h) narrate a series of major events located on a timeline and use finite, dynamic verbs in the past tense. Conversely, backgrounded portions, providing secondary information, involve a variety of structures: the non-finite verb forms in (b), (e), and (i) a stative clause in (c) and a subordinate clause in (d).

7.2.9 Viewpoint

An essential feature of stories (whether fact or fiction) is that they represent the speech, thoughts, attitudes, and emotions of participants and of authors/storytellers. In processing narrative discourse, listeners/readers construct conceptualizations of the ways these different viewpoints are connected into a meaningful network and connect it to their own point of view, thus adding a further filter that changes the meaning of the story. This

depends, for example, on whether the listener identifies herself with one or the other event participant (a killer or a victim, for instance).

The study of the complexities of viewpoint in narrative discourse may provide an especially interesting window on core characteristics of human cognition, while theories of social cognition and its evolution may shed light on the delight that humans universally take in story-telling and the role of viewpoint in it.(Vandelanotte 2017)

7.2.10 Evaluation and stance

The notion of 'evaluation' is meant to cover whatever emotions, feelings, and attitudes are associated with perceptual experiences or inner mental activities (memories, previous experiences). Among the variable properties of consciousness singled out by Chafe (W.. Chafe 1994) is that an experience may be derived from the immediate environment or be displaced (in time), it may be a product of remembering or imagining, or involve judgments as to whether the content of consciousness is factual or fictional.

This is reflected in languages in different ways. Grammars express, for example, the *realis/irrealis* distinction ('realis' is concerned with real (and necessary) events, such as 'Water boils at 100 degrees' whereas 'irrealis' expresses what is considered hypothetical, conditional, possible or imaginary, as in 'I heard he had a mistress'. This dichotomy may be associated with verb aspect/tense (e.g., subjunctive, future tense, hypothetical clauses and conditionals are all close to the irrealis end of the scale), or with the presence of so called evidentials, i.e., means of expressing the source of information on which the speaker bases what he is saying (he might have witnessed something first-hand, or just report hearsay or gossip). Some languages have specific morphemes to express that. Others, such as English, encode it through lexical expressions such as adverbials (visibly, reportedly, apparently, allegedly) or by verbs in main-clause or parenthetical types of construction ('I see', 'I hear', 'the witness said'), propositional attitude verbs ('I guess', 'I suppose', 'I realize', 'I feel', 'I imagine'), speech act verbs ('I swear', 'I promise', 'I predict'), mood (declarative versus interrogative), tense (present versus conditional).

7.2.11 Speech acts

The concept of speech act is crucial in pragmatics: the core idea is that utterances are not mere meaning-bearers. Rather, they perform actions, thus affecting reality and interaction dynamics between people. In other words, an utterance not only has a meaning, but it has also a specific use ('force'). Utterances often have non-verbal counterparts (cf. waving to saying hello, bidding at auction by hand or voice) as well as real-world consequences just like non- verbal actions (a 1,000 dollar bid at an auction commits you to paying (Levinson 2017). Examples of explicit and implicit speech acts are:

(15) I'll come to your event. (Promise)

(16) There's a bull in the arena! (Warning)

And some more examples from the crime television series The Mentalist, season 5, episode 18 with the context that Jane is at the hospital and gets a phone call from Lisbon:

(17) Lisbon: Don't tell me you are at the hospital.
Jane: I am not at the hospital."

In (17), Jane deliberately misinterprets the meaning evoked by Lisbon. Her utterance was a rhetorical question, while Jane replies as if it were a speech act (more specifically an order).

7.2.12 Pragmatics and sociolinguistics: some applications

The attitude, stance, and purpose of the speaker/narrator is evident if one considers how narratives are built. Labov (Labov 2006) highlights how a narrative of an event involves a prior cognitive decision that a given event is reportable: an event narration often consists of chains of events that are linked causally each to the following one. Comparison of such event chains with the sequence of narrative clauses actually produced will help understand how the narrator re-organizes and transforms the events of real time in the finished narrative. The way a narrative is told has also important implications in various sociolinguistic realms. As noticed earlier by Hymes and Cazden (Hymes and Cazden 1980), choice between emotionally marked language registers versus more objective, detached expressions and linguistic constructions affects the perception of the truthfulness of what is being said (in their study, in university classroom discussions). Similarly, Blommaert (Blommaert 2001) offers an interesting case study on narrative inequality: it discusses African asylum seekers stories in Belgium, highlighting the importance of the narrative structure and characteristics towards a successful application in the asylum procedure. In short, how you frame your narrative (lexical, syntactic, register choices etc.) has a crucial role in truth assessment and evaluations in general that the reader/listener, consciously or not, carries on.

7.3 Conclusions

As shown throughout the paper, pragmatics is a critical point to fully understand how a textual or spoken narration is built and comprehended: specifically, the role of context, contextualization and anchoring to previous knowledge, or world knowledge as a whole, is a central issue. Also the means through which the text introduces referents (event participants and entities) and the ways it refers back to those referents are important issues, involving different referential cues and anaphoric means. Finally, languages have means to express the attitude, purpose, and stance of the speaker so that the listener can reconstruct also those aspects of a narrative.

.

Chapter 8

Conclusions
Luc Steels

Abstract

This concluding chapter summarizes some of the main insights developed in the papers contained in this volume and outlines some of the key research questions that require significant breakthroughs if we want to build AI systems that understand.

Keywords

Meaning, understanding, human-centric AI, meaningful AI, narratives.

8.1 Main insights

The introductory chapter to this volume (Chapter 1. Conceptual Foundations of Human-centric AI) started from the **distinction between reactive and deliberative AI**, called system 1 and system 2 by Daniel Kahneman in his well known book 'Thinking Fast and Slow'. Reactive intelligence is based on stimulus-response patterns that give immediate solutions to problems and is therefore fast. Today it is studied in AI mainly through Machine Learning, for example through re-inforcement learning or supervised deep learning. Deliberative AI is based on rich models and reasoning. It is consequently slower. It is based on ontologies, very large knowledge bases, fine-grained language processing and reasoning. Reactive intelligence is needed for sensory-motor intelligence, for making quick decisions, or getting a good 'intuitive' guess faced with multiple hypotheses and a potentially exploding search space. Deliberative intelligence is needed for problem solving, dealing with unforeseen situations, human-understandable explanations, validation and tutoring.

Even though reactive intelligence has proven to lead to surprisingly high performance on many tasks, there is a growing consensus in AI research that **a combination of reactive and deliberative intelligence is needed**. This combination goes under the heading of 'integrated' or 'integrative AI', or also 'composite' or 'hybrid AI' (although hybrid AI is also used for AI where humans work intimately together with artificial systems). Such a combination is particularly relevant for human-centric AI, in order to achieve properties such as explainability, robustness, verifiability and ethical and moral acceptability.

Deliberative intelligence relies on the construction of a 'rich' model before deliberative intelligence can start its work. For example, planning and executing the necessary actions to cook a recipe requires a detailed model of the ingredients and the cooking activities, solving an algebraic problem phrased in natural language requires first that the problem description is formalized into a set of equations which an equation solver can then solve. In a lot of AI work these models are supplied by design. For example, a planner has to be given precise inputs, goals and a series of possible actions before it can start to do the actual planning.

We have defined **understanding as the process of constructing the rich model on which deliberative intelligence relies.** Understanding must typically handle various kinds of inputs: text, image, sound, embodied action. It has to ground the model both in the perceptions or data about the world *and* in the semantic knowledge and past experiences of the agent. It calls upon a variety of knowledge sources:

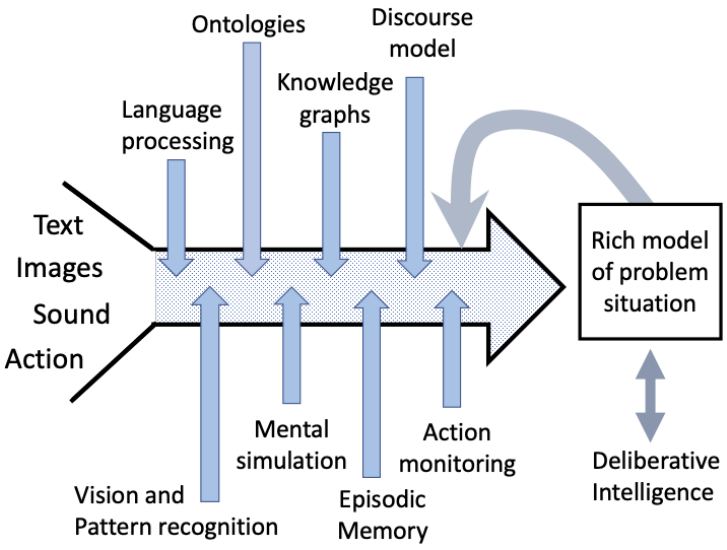


Figure 8.1: Understanding calls upon a variety of knowledge sources and uses itself the partially constructed rich model. The knowledge sources use both reactive and deliberative AI.

- Language processing for decoding input texts.
- Ontologies that define relevant frames with slots, defaults and constraints on slot fillers.

- Knowledge graphs that constitute a semantic memory of general and specific facts.
- Discourse models for tracking the flow of attention in dialogs.
- Action monitoring for perceiving and interpreting actions.
- Episodic memory that contains a memory of models made for past situations.
- Mental simulation that either in a quantitative or qualitative way simulates future world states.
- Vision and pattern recognition for signal processing, segmentation and pattern recognition of images or sound.

8.2 Prior Art

Understanding is very hard because it requires significant advances in all areas of AI and integration of AI subfields that have often working on their own. Understanding can therefore be considered the central unsolved hard problem of AI. It is also hard because in real world situations inputs are typically sparse, fragmentary, ambiguous, underspecified, uncertain, occasionally contradictory and possibly deliberately biased (for example because the producer of inputs is trying to deceive or manipulate). Often different solution paths have to be considered with the risk of exploding combinatorial complexity and the knowledge sources available are themselves sparse, fragmentary, uncertain and logically incoherent. Furthermore understanding is hard because of the hermeneutic paradox: to understand the whole you need to understand the parts but to understand the parts you need to understand the whole. This calls for other control structures than a simple linear flow of processes organised in a pipeline.

Understanding is the process of building rich models. These models will be in line with what has been practiced now for decades in the field of knowledge representation, knowledge-based systems, fine-grained language processing and semantic web technologies, i.e. they will consist of a very large network of concepts and relations between concepts, often organised in terms of frames and augmented with methods for inference and meta-information to handle uncertainty, defaults, inconsistencies, etc.

In addition, the investigations reported in this volume point to important characteristics of human-made rich models which have so far not been considered much in AI but are in fact essential for domains where social issues are at stake. The humanistic and social science disciplines studying these domains (sociology, economics, linguistics, history, semiotics) characterize these models as narratives and many concrete examples were given in chapters 2-6. Narratives form a continuum from scientific or realistic models of the world where veracity is high, rationality is high and rhetoricity is low, to fictional narratives where veracity is low, rationality is low and rhetoricity is high. The purpose of fictional narratives is not to get to the truth by verified facts and logical arguments but to be compatible with

shared values and past experiences and convince the reader from the viewpoints held by the author. This point has been illustrated repeatedly by the different domains discussed in the papers contained in this volume.

There was already a flurry of research activity in AI on narratives between the mid nineteen-sixties and late nineteen-seventies. One of the first ideas that came out of this research is the notion of a semantic network, as first proposed by Quillian(Quillian 1968), which blossomed into the knowledge graphs that we have today.

Another notion is that of a schema or frame, as first proposed by Minsky (Minsky 1975), or script, as proposed and worked out by Schank and colleagues (Schank and Abelson 1977). A frame is a set of questions to be asked about a particular situation with constraints, defaults and strategies to find answers. The classical example is a restaurant frame that structures the experience of entering a restaurant.

Frames and scripts have underpinned a large number of earlier experimental AI systems for story understanding and story generation and for expert systems not based on rules but on solving cases by analogy with earlier cases. They also lead to some remarkable early demonstrations of language understanding capacities in the 1970s, such as Winograd's SHRDLU system, dubbed a system that could understand natural language,(Winograd 1976) and the HEARSAY speech understanding system that worked on spoken language.(Reddy et al. 1974) Later in the 1980s we also saw the first attempts to build encyclopedic knowledge bases, of which CYC (Lenat 1995) was the most important representative.

Schank also triggered the first AI research into dynamic semantic and episodic memory(Schank 1990) which developed further in memory-based or case-based AI applications.(Kolodner 1992)

Interest in narratives waned somewhat in the late 1990s as the attention of AI shifted to behavior-based robotics, neural networks, and the semantic web. Moreover it was realized that incorporating narratives as the core of AI was going to be a daunting task. The broad humanistic scope that characterized AI in its first decades began to shrink with a focus on narrowly defined problems with measurable performance under economic pressures to come up with exploitable results. But the past decade we have seen renewed attention into narratives (Mateas and Sengers 2003), (Winston 2011), (Finlayson 2013), (Riedl 2016), (Gervás et al. 2019), (Meghini, Bartalesi, and Metilli 2021), now incorporating also data-driven AI methods to learn aspects of narrative intelligence, such as the reconstruction of timelines or characters, from corpora. Still, it is early days and many problems remain unsolved.

8.3 Key issues for narrative-based AI

What are some of the priority issues we should focus on today? We restrict ourselves here to four directions in which further work is needed.

1. *Improving knowledge sources*: Despite the fact that reactive AI has made many steps forward and many more semantic and language resources are now available, all of this still falls far short of what is needed for understanding. Given the current high investment in AI (particularly in reactive AI) we can expect significant advances in the coming decade. Still, key research is needed that focuses on narratives and narrations. For example, in the case of language processing much more effort is required to engage with the pragmatic aspects of language (as discussed in chapter 7), or in the case of knowledge graphs we need more effort to traverse graphs for finding new links (as discussed in chapter 4 for clinical narratives). (I. Tiddi, M. D'Aquin, and E. Motta 2014)

2. *Computational representation of narratives*: It is obvious that we need a powerful data structure that captures all the information about a narrative including its narration. This data structure must represent the facts in the fabula, the way the facts are selected, framed and organized in a plot, and the intermediary structures that are being built to produce or comprehend a narration. This data structure should take the form of a narrative network, similar to a knowledge graph, but with much more information and meta-information.

A narrative network is transient in the sense that it grows and changes as more information comes in during the understanding process. It has to be able to represent different hypotheses given the ambiguity and under-specification of inputs and the uncertainties associated with facts or inferences. It should support the exploration of multiple hypotheses and adapt to progressive insight or changes, both in the inputs and in the available knowledge sources.

There have been plenty of experiments and technical advances in the design and implementation of narrative networks, for example for dealing with uncertainty or dealing with the exploration of multiple hypotheses, but a big challenge remains: to create a comprehensive design and, if possible, standardize it so that many people can collaborate as has happened with ontologies, grammars and knowledge graphs. Work on narrative annotation tools and formal ontologies already goes in this direction.(Finlayson 2011), (Meghini, Bartalesi, and Metilli 2021), (Porzel 2021)

3. *Cognitive architecture* Many different knowledge sources contribute to the build up of a narrative network and we certainly need to go beyond a strict pipeline model that dominates data-driven AI towards a flexible architecture in which different knowledge sources can contribute at any time, either because new information has become available, so that a knowledge source can make a useful contribution, or because they are called in a top-down manner to expand areas of the narrative network. In past AI work, this kind of flexibility has been approached with blackboard architectures based on the metaphor of a blackboard on which various knowledge sources can read and write.(Englemore and T. Morgan 1988) In the case of understanding the blackboard contains the transient narrative network and the knowledge sources include sensory input, language input, mental simulation, semantic and episodic memory, and more. Blackboard architectures have resonated with models of consciousness in neuroscience, specifically the Global Neural Workspace

Model by Stanislas DeHaene and colleagues.(De Haene, Changeux, and Naccache 2011)

A blackboard-like architecture is a first step but we will need much more:

- First of all, we need ways in which the understanding system can *measure progress*: (i) how far ambiguity, incompleteness, under-specification and incompleteness is being reduced, (ii) how information has become integrated, meaning in how far network fragments could be connected to each other and how connections between different levels of the narrative (the fabula, plot, and narration) could be established, and (iii) how far the distance towards a satisfactory narrative closure could be bridged. Narrative closure happens when the key questions for which the narrative is being developed have been answered.

- Second, these measures should be input to an *attention mechanism* that decides where further effort should go: Should more resources be applied for analyzing and interpreting certain inputs? Which nodes in the network are to be expanded preferentially? What additional inputs might profitably be sought? It is unavoidable that the degree of understanding of a narration will sometimes decrease because new input enters that is not compatible with what was seen before or new facts are supplied by semantic memory that create a cognitive dissonance with the narrative network built so far.

- Third, there needs to be a *meta-level* that plays three crucial roles: (a) It should monitor progress and avoid catastrophic failure by catching fail states in components (in the same way operating system catches errors in applications and possibly repairs them to avoid that the whole system collapses). (b) It should govern learning processes, such as the process by which a narrative is stored in episodic memory to deal with similar situations in the future. (c) It should include a value system that monitors decision-making (and subsequent real-world action) to make it compatible with a moral framework that is compatible with human values in order to realize *value-aware AI*.

3. *Other paths to learning:* Current AI applications focus on a narrow set of learning mechanisms but we need to become adventurous (again) and explore other learning strategies. We just list three paths for further exploration:

(a) It was argued in the section on meaning that the distinctions (categories/concepts) used in AI systems should be meaningful with respect to tasks and contexts and that an exclusive focus on prediction does not yield the kind of distinctions that humans find natural or fit with other tasks than prediction. So we need to find new frameworks for learning. In the case of language, one such framework are language games.(Luc Steels 1998) Language games are played by two agents selected out of a population of agents, possibly including humans. They have a communicative task, such as drawing attention to an object in their shared world-setting or giving instructions to carry out a certain action and use language to do so. The agents start without a shared language system or a shared ontology of distinctions and have to build that up as part of becoming successful in

playing games. This approach has now been followed for dozens of domains (color, space, time, action, etc.)(L. Steels 2012b) and for many different aspects of language.(L. Steels 2016) It shows that categories need not be formed by induction over a large dataset only but can be learned in an incremental fashion and relevant for the task of communication.

(b) Another possible path are alternative learning mechanisms. The coupling of reactive intelligence to deliberative intelligence provides new avenues. For example, we could explore more constructivist (or abductive) approaches to learning domain knowledge for deliberative intelligence. A constructivist approach suggests that the learner is able to partially solve problems and then examine what domain knowledge is missing in order to complete the solution. Or the learner solves a problem but gets feedback whether the derived solution is adequate. If not, constructivist learning analyses in which way domain knowledge can be changed and repair it.(L. Steels 2004)

(c) Yet another path is a much more tight coordination between reasoning and learning, as happened in earlier explanation-based learning approaches. Learning is not just in terms of compiling strategies to cut search (as in the alphaGo experiments) but of hypothesizing new narrative structures that can be used for improved deductions. Examples of this were discussed in chapter 4 on clinical narratives.

8.4 Conclusions

The main outcomes of the studies reported in this volume are (i) a clear definition of understanding that can be the basis of computational experiments, (ii) examples from many cognitive, social and humanistic sciences of the nature of human models, namely as narratives, (iii) preliminary insights on how the understanding process can be organized.

Much remains to be discovered but the path for exploring understanding, one of the hardest problems in AI, is now much clearer.

.

Bibliography

Abeysinghe, Sudeepa (2015). "Vaccine narratives and public health: Investigating criticisms of H1N1 pandemic vaccination". In: *PLoS currents* 7.

Ahern, Kenneth R and Denis Sosyura (2014). "Who writes the news? Corporate press releases during merger negotiations". In: *The Journal of Finance* 69.1, pp. 241–291.

Akerlof, George A (2020). "Sins of Omission and the Practice of Economics". In: *Journal of Economic Literature* 58.2, pp. 405–18.

Akerlof, George A and Dennis J Snower (2016). "Bread and bullets". In: *Journal of Economic Behavior & Organization* 126, pp. 58–71.

Akerlof, Robert, Niko Matouschek, and Luis Rayo (2020). "Stories at work". In: *AEA Papers and Proceedings*. Vol. 110, pp. 199–204.

Anderson, Tory S (2015). "From episodic memory to narrative in a cognitive architecture". In: *6th Workshop on Computational Models of Narrative (CMN 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Andrade-Castellanos, Carlos A. et al. (2016). "Subcutaneous rapid-acting insulin analogues for diabetic ketoacidosis". In: *Cochrane Database of Systematic Reviews* 2016.1. ISSN: 1469493X. DOI: 10.1002/14651858.CD011281.pub2.

Andrikopoulos, Andreas (2013). "Financial economics: objects and methods of science". In: *Cambridge Journal of Economics* 37.1, pp. 35–55.

Antoniou, G. and F. Van Harmelen (2008). *A semantic Web Primer*. Cambridge Ma: The MIT Presss.

"Anytime bottom-up rule learning for knowledge graph completion" (2019). In: *IJCAI International Joint Conference on Artificial Intelligence* 2019-Augus, pp. 3137–3143. ISSN: 10450823. DOI: 10.24963/ijcai.2019/435.

Apesos, A. (2010). "The painter as Evangelist in Caravaggio's Taking of Christ ." In: *Aurora* XI.

Arthur, W Brian (2014). *Complexity and the Economy*. Oxford University Press.

Ashenfelter, Orley (2012). "Economic History or History of Economics? Grand Pursuit: The Story of Economic Genius: Review Essay". In: *Journal of Economic Literature* 50.1, pp. 96–102.

Atran, S. (2003). "Genesis of suicide terrorism." In: *Science* 299(5612), pp. 1534–9.

— (2010). *Talking to the enemy: Faith, brotherhood, and the (un) making of terrorists*. London: Penguin books.

Atran, S. (2021). "Psychology of Transnational Terrorism and Extreme Political Conflict." In: *Annual Review of Psychology* 72, pp. 471–501.

Atran, S. and R. Axelrod (2008). "Reframing sacred values: In theory." In: *Negotiation Journal* 24(3), pp. 221–46.

Atran, S. and J. Henrich (2010). "The Evolution of Religion: How Cognitive By-Products, Adaptive Learning Heuristics, Ritual Displays, and Group Competition Generate Deep Commitments to Prosocial Religions." In: *Biological Theory* 5(10), pp. 18–30.

Bakal, Gokhan et al. (2018). "Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations". In: *Journal of Biomedical Informatics* 82.January, pp. 189–199. ISSN: 15320464. DOI: 10.1016/j.jbi.2018.05.003.

Bal, M. and C. Van Boheemen (1997). *Narratology: Introduction to the theory of narrative.* Toronto, Canada: University of Toronto Press.

Baltag, Alexandru and Sonja Smets (2013). "Protocols for belief merge: Reaching agreement via communication". In: *Logic Journal of the IGPL* 21.3, pp. 468–487.

Barnes, J. (2004). *The pedant in the kitchen.* London: Atlantic Books.

Baron, J. and M. Spranca (1997). "Protected Values." In: *Virology* 70(1), pp. 1–16.

Bauer, Nico et al. (2017). "Shared socio-economic pathways of the energy sector–quantifying the narratives". In: *Global Environmental Change* 42, pp. 316–330.

Beetz, M. et al. (2012). "Cognition-Enabled Autonomous Robot Control for the Realization of Home Chore Task Intelligence". In: *Proceedings of the IEEE* 100.8, pp. 2454–2471.

Bender, E. and Koller A. (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings Annual Meeting of Association for Computational Linguistics Conference.* Association for Computational Linguistics, 5185–A5198.

Béné, Christophe et al. (2019). "When food systems meet sustainability–Current narratives and implications for actions". In: *World Development* 113, pp. 116–130.

Berry, Dominic J. (2021). "Narrative and Epistemic Positioning: The Case of the Dandelion Pilot". In: *Engineering and Philosophy: Reimagining Technology and Social Progress.* Ed. by Zachary Pirtle, David Tomblin, and Guru Madhavan. Cham: Springer, pp. 123–139. DOI: https://doi.org/10.1007/978-3-030-70099-7_6.

Beuls, K. (2017). "A computational construction grammar approach to semantic frame extraction." In: *Constructions and Frames* 9(2).

Bijker, Wiebe E, Roland Bal, and Ruud Hendriks (2009). *The paradox of scientific authority: The role of scientific advice in democracies.* MIT press.

Blommaert, J. (2001). "Investigating Narrative Inequality: African Asylum Seekers' Stories in Belgium." In: *Discourse & Society* 12(4), pp. 413–449.

Blomqvist, Eva, Marjan Alirezaie, and Marina Santini (2020). "Towards Causal Knowledge Graphs-Position Paper". In: *KDH@ ECAI.*

Boer, V. de et al. (2013). "Amsterdam museum linked open data." In: *Semantic Web* 4(3), pp. 247–243.

Boswell, John (2013). "Why and how narrative matters in deliberative systems". In: *Political Studies* 61.3, pp. 620–636.

Bragge, P. (2010). "Asking good clinical research questions and choosing the right study design". In: *Injury* 41.SUPPL.1, S3. ISSN: 00201383. DOI: 10.1016/j.injury.2010.04.016.

Brinton, Crane (1965). *Anatomy of a Revolution*. First ed. 1938, revised ed. New York: Vintage Books.

Brown, Mary Helen (1986). "Sense making and narrative forms: Reality construction in organizations". In: *Organizational Communication; Emerging Perspectives*, pp. 71–84.

Bruner, J. (1991). "The Narrative Construction of Reality." In: *Critical Inquiry* 18(1), pp. 1–21.

Bruner, Jerome (1990). *Acts of meaning*. Harvard University Press.

Bruner, Jerome Seymour (2003). *Making stories: Law, literature, life*. Harvard University Press.

Chafe, W. (1974). "Language and consciousness." In: *Language* 50, pp. 111–133.

— (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex.

— (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago, IL: The University of Chicago Press.

Chan, Carol (2014). "Gendered morality and development narratives: the case of female labor migration from Indonesia". In: *Sustainability* 6.10, pp. 6949–6972.

Charon, R. et al. (2017). *The principles and practice of narrative medicine*. Oxford UK: Oxford University Press.

Charters, Erica and Kristin Heitman (2021). "How Epidemics End". In: *Centaurus* 63.1, pp. 210–224. DOI: https://doi.org/10.1111/1600-0498.12370.

Chazelas, Jean (1968). "La Suppression de la Gabelle du Sel en 1945". In: *Le Rôle du Sel dans l'Histoire*. Ed. by Michel Mollat. Paris: Presses Universitaires de France, pp. 263–265.

Cherif, Reda, Marc Engher, and Fuad Hasanov (2020). "Crouching Beliefs, Hidden Biases: The Rise and Fall of Growth Narratives". In.

Chomsky, N. (1956). "Three models for the description of language.." In: *IRE Transactions on Information Theory* 2(3), pp. 113–124.

Copeland, Morris A (1931). "Economic theory and the natural science point of view". In: *The American Economic Review*, pp. 67–79.

Copland, A. (1939). *What to listen for in music*. London: Penguin.

Costola, Michele, Matteo Iacopini, and Carlo R.M.A. Santagiustina (2021a). "On the" mementum" of Meme Stocks". In: *arXiv preprint arXiv:2106.03691*.

— (2021b). "Google search volumes and the financial markets during the COVID-19 outbreak". In: *Finance research letters* 42, p. 101884.

Cowan, N. (2008). "What are the differences between long-term, short-term, and working memory?" In: *Progress in Brain Research* 169, pp. 323–338.

Crouch, Colin (2011). *The strange non-death of neo-liberalism*. Polity.

Currie, Adrian Mitchell (2014). "Narratives, Mechanisms and Progress in Historical Science". In: *Synthese* 191, pp. 1163–1183. DOI: https://doi.org/10.1007/s11229-013-0317-x.

Das, Rajarshi et al. (2018). "Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning". In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. eprint: 1711.05851.

De Cao, Nicola and Thomas Kipf (2018). "MolGAN: An implicit generative model for small molecular graphs". In: *arXiv preprint arXiv:1805.11973*.

De Haene, S., J-P. Changeux, and L. Naccache (2011). "The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications." In: *Research and Perspectives in Neurosciences* 18, pp. 55–84.

Decuyper, J-C, D. Keymeulen, and L. Steels (1995). "A Hybrid Architecture for Modeling Liquid Behavior." In: *Diagrammatic Reasoning*. Ed. by J. Glasgow, N. Narayanan, and B. Chandrasekaran. Menlo Park: AAAI Press.

Dehghani, M. et al. (2010). "Sacred values and conflict over Iran's nuclear program." In: *Judgement and Decision Making* 5, pp. 540–546.

Devlin, Jacob et al. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL-HLT (1)*. Association for Computational Linguistics, pp. 4171–4186.

DiMaggio, Paul et al. (2001). "Social implications of the Internet". In: *Annual review of sociology* 27.1, pp. 307–336.

Dolan, Paul et al. (2012). "Influencing behaviour: The mindspace way". In: *Journal of Economic Psychology* 33.1, pp. 264–277.

Donald, M. (2006). "Art and Cognitive Evolution." In: ed. by M. Turner. Oxford, UK: Oxford University Press.

Douglas, K., R. Sutton, and A. Cichocka (2017). "The psychology of conspiracy theories." In: *Current Directions in Psychological Science* 26(6).

Drancé, Martin et al. (2021). "Neuro-symbolic XAI for Computational Drug Repurposing". In: 2.Ic3k, pp. 220–225. DOI: 10.5220/0010714100003064.

Du Bois, J. (1987). "The discourse basis of ergativity." In: *Language* 64, pp. 805–855.

Duchan, Judith F, Gail A Bruder, and Lynne E Hewitt (2012). *Deixis in narrative: A cognitive science perspective*. Psychology Press.

Dumez, Hervé and Alain Jeunemaitre (2006). "Reviving narratives in economics and management: towards an integrated perspective of modelling, statistical inference and narratives". In: *European Management Review* 3.1, pp. 32–43.

Eco, U. (1975). *Trattato di semiotica generale*. Milano: La nave di Teseo.

— (1979). *The Role of the Reader: Explorations in the Semiotics of Texts*. Bloomington In: Indiana University Press.

Eichengreen, Barry (2012). "Economic history and economic policy". In: *The Journal of Economic History* 72.2, pp. 289–307.

Ekelund, Robert Burton et al. (1999). *Secret origins of modern microeconomics: Dupuit and the engineers*. University of Chicago Press.

Elster, Jon (1979). "Ulysses and the sirens: Studies in rationality and irrationality". In.
Englemore, R. and T. Morgan (1988). *Blackboard Systems.* New York: Addison-Wesley.
Erlich, Victor (1973). "Russian Formalism". In: *Journal of the History of Ideas* 34.4, pp. 627–638.
Erman, L. et al. (1980). "Knowledge to Resolve Uncertainty." In: *Computing Surveys* 12(2), pp. 213–253.
Evans, George W and Seppo Honkapohja (2012). *Learning and expectations in macroeconomics*. Princeton University Press.
Evans, Vyvyan (2009). *How words mean: Lexical concepts, cognitive models, and meaning construction*. Oxford University Press on Demand.
Feigenbaum, E. (1977). "The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering". In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. IJCAI, pp. 1014–1029.
Feinberg, Joshua et al. (2017). "Nutrition support in hospitalised adults at nutritional risk". In: *Cochrane Database of Systematic Reviews* 2017.5. ISSN: 1469493X. DOI: 10.1002/14651858.CD011598.pub2.
Ferguson-Cradler, Gregory (2021). "Narrative and computational text analysis in business and economic history". In: *Scandinavian Economic History Review*, pp. 1–25.
Ferrara, Federico Maria et al. (2021). "Exports vs. investment: How political discourse shapes popular support for external imbalances". In: *Socio-Economic Review*.
Fillmore, C. (1976). "Frame semantics and the nature of language." In: *Annals of the New York Academy of Sciences* 280(1), pp. 20–32.
Finlayson, M. (2011). "The story workbench: An extensible semi-automatic text annotation tool." In: *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. WS-11-18.
— (2013). "A Survey of Corpora in Computational and Cognitive Narrative Science." In: *Sprache und Datenverarbeitung (International Journal for Language Data Processing)* 37(1-2), pp. 113–141.
Fisher, W. (1987). "Human communication as narration: toward a philosophy of reason, value, and action." In: *Studies in rhetoric communication* 40(2).
Foley, W. and R. Van Valin (1984). *Functional Syntax and Universal Grammar.* Cambridge, UK: Cambridge University Press.
Forbus, K. (1988). "Qualitative Physics: Past, Present and Future." In: *Exploring Artificial Intelligence.* Amsterdam: Elsevier, pp. 239–296.
Forbus, K. and J. De Kleer (1993). *Building Problem Solvers.* Cambridge Ma: The MIT Press.
Fourcade, Marion (2009). *Economists and societies*. Princeton University Press.
Fox, Edward J and Stephen J Hoch (2005). "Cherry-picking". In: *Journal of Marketing* 69.1, pp. 46–62.
Freedman, Lawrence (2015). "The possibilities and limits of strategic narratives". In: *Strategic Narratives, Public Opinion and War*. Routledge, pp. 45–64.
Fu, Gang et al. (2016). "Predicting drug target interactions using meta-path-based semantic network analysis". In: *BMC Bioinformatics* 17.1, pp. 1–10. ISSN: 14712105.

DOI: 10.1186/s12859-016-1005-x. URL: http://dx.doi.org/10.1186/s12859-016-1005-x.

Gadamer, H-G.. (1975). "Hermeneutics and Social Science." In: *Cultural Hermeneutics* 2(4), pp. 207–316.

Gatrell, Jay D and Neil Reid (2017). "Placing economic development narratives into emerging economic spaces: Project Jeep, 1996-1997". In: *New economic spaces: New economic geographies*. Routledge, pp. 108–118.

Gaur, Monika and Ravi Kant (2021). "Can Narrative Economics Justify Economic Fluctuations and Inequality? An Approach from Micro to Macro Perspective". In: *Theoretical Economics Letters* 11.01, p. 1.

Gervás, Pablo et al. (Jan. 2019). "The Long Path to Narrative Generation". In: *IBM Journal of Research & Development* 63.1, pp. 1–8.

Ginges, J. (2007). "Sacred bounds on rational resolution of violent political conflict." In: *Proceedings of the National Academy of Sciences of the United States of America.* 104(18), pp. 7357–60.

Ginges, J. et al. (2011). "Psychology out of the laboratory: the challenge of violent extremism." In: *The American Psychologist.* 66(6), pp. 507–519.

Givón, T. (1984). *Topic Continuity in Discourse: a quantitative cross-language study.* Amsterdam: John Benjamins Publishing.

Godechot, Olivier (2011). "How Did the Neoclassical Paradigm Conquer a Multi-disciplinary Research Institution?. Economists at the EHESS from 1948 to 2005". In: *Revue de la régulation. Capitalisme, institutions, pouvoirs* 10.

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience.* New York: Harper and Row.

Goudge, Thomas A. (1958). In: *The British Journal for the Philosophy of Science* 9.35, pp. 194–202.

Gu, Yulong and Paolo Missier (n.d.). "Towards Learning Instantiated Logical Rules from Knowledge Graphs". In: (). eprint: arXiv:2003.06071v2.

Guest, Ivor, ed. (1960). *La Fille Mal Gardée.* London: Dancing Times.

Haidt, J. (2008). *The Moral Roots of Liberals and Conservatives.*

Hall, Sarah (2006). "What counts? Exploring the production of quantitative financial narratives in London's corporate finance industry". In: *Journal of Economic Geography* 6.5, pp. 661–678.

Hamid, N. et al. (2019). "Neuroimaging 'will to fight' for sacred values: An empirical case study with supporters of an Al Qaeda associate." In: *Royal Society Open Science* 6(6).

Hanzlick, Randy L., ed. (1997). *Cause-of-Death Statements and Certification of Natural and Unnatural Deaths. Protocol and Options.* Northfield IL: College of American Pathologists.

Harris, Seymour E et al. (1959). "Round table on the organization and financing of economic research". In: *The American Economic Review* 49.2, pp. 559–580.

Hartig, Olaf and M Tamer Özsu (2016). "Walking without a map: Ranking-based traversal for querying linked data". In: *International Semantic Web Conference.* Springer, pp. 305–324.

Heath, Robert L, Jaesub Lee, and Laura L Lemon (2019). "Narratives of risk communication: Nudging community residents to shelter-in-place". In: *Public Relations Review* 45.1, pp. 128–137.

Herlihy, David and Christiane Klapisch-Zuber (1985). *Tuscans and their Families: A Study of the Florentine Catasto of 1427*. New Haven: Yale University Press.

Herrmann-Pillath, Carsten (2008). "The naturalistic turn in economics: implications for the theory of finance". In: *Available at SSRN 1289588*.

Hewitt, Elizabeth (June 2020). *Speculative Fictions: Explaining the Economy in the Early United States*. en. Oxford University Press. ISBN: 978-0-19-260299-2.

Himmelstein, Daniel Scott et al. (2017). "Systematic integration of biomedical knowledge prioritizes drugs for repurposing". In: *Elife* 6, e26726.

Hogan, Aidan et al. (2020). *Knowledge Graphs*. arXiv: 2003.02320 [cs.AI].

Holzinger, Andreas et al. (2019). "Causability and explainability of artificial intelligence in medicine". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4, pp. 1–13. ISSN: 19424795. DOI: 10.1002/widm.1312.

Hullman, Jessica and Nick Diakopoulos (2011). "Visualization rhetoric: Framing effects in narrative visualization". In: *IEEE transactions on visualization and computer graphics* 17.12, pp. 2231–2240.

Hume, Margee and Michael Mills (2013). "Uncovering Victoria's Secret: Exploring women's luxury perceptions of intimate apparel and purchasing behaviour". In: *Journal of Fashion Marketing and Management: An International Journal*.

Hymes, C. and C. Cazden (1980). "Narrative Thinking and Storytelling Rights: A Folklorist's Clue to tique of Education." In: *Language and Education: Ethnolinguistic Essays*. Ed. by C. Hymes. New York, NY: Center for Applied Linguistics, pp. 126–138.

Iacopini, Matteo and Carlo R.M.A. Santagiustina (2021). "Filtering the intensity of public concern from social media count data with jumps". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184.4, pp. 1283–1302. DOI: https://doi.org/10.1111/rssa.12704.

Irvine, L., J. Pierce, and R. Zussman (ed) (2019). *Narrative Sociology*. Nashville, Te: Vanderbilt University Press.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Strauss and Giroux.

Kahneman, Daniel et al. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Keen, Steve (2001). *Debunking economics: The naked emperor of the social sciences*. Zed Books.

Keusch, Thomas, Laury HH Bollen, and Harold FD Hassink (2012). "Self-serving bias in annual report narratives: An empirical analysis of the impact of economic crises". In: *European Accounting Review* 21.3, pp. 623–648.

Keven, Nazim (2016). "Events, narratives and memory". In: *Synthese* 193.8, pp. 2497–2517.

Kirkwood, William G (1992). "Narrative and the rhetoric of possibility". In: *Communications Monographs* 59.1, pp. 30–47.

Kitromilides, Yiannis (2013). "Stories, fables, parables, and myths: Greece and the Euro crisis, toward a new narrative". In: *Journal of Economic Issues* 47.3, pp. 623–638.

Klic, L. (2021). "Linked Open Images. Visual similarity for the semantic web." In: *Semantic Web* 1, pp. 1–15.

Kolodner, J. (1992). "An introduction to case-based reasoning.." In: *Artificial Intelligence Review* 6, pp. 3–34.

Kroll, H., D. Nagel, and W. Tilo-Balke (2020). "Modeling Narrative Structures in Logical Overlays on Top of Knowledge Repositories". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12400 LNCS, pp. 250–260. ISSN: 16113349. DOI: 10.1007/978-3-030-62522-1_18.

Kroll, Hermann, Denis Nagel, and Wolf Tilo Balke (2020). "Modeling Narrative Structures in Logical Overlays on Top of Knowledge Repositories". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12400 LNCS, pp. 250–260. ISSN: 16113349. DOI: 10.1007/978-3-030-62522-1_18.

Kroll, Hermann, Denis Nagel, Morris Kunz, et al. (2021). "Demonstrating Narrative Bindings: Linking Discourses to Knowledge Repositories." In: *Text2Story@ ECIR*, pp. 57–63.

Kroner, Alexander et al. (2019). "Contextual Encoder-Decoder Network for Visual Saliency Prediction". In: *CoRR* abs/1902.06634. URL: http://arxiv.org/abs/1902.06634.

Kuempel, M., C. Mueller, and M. Beetz (2021). "Semantic Digital Twins for Retail Logistics". In: *Dynamics in Logistics: Twenty-Five Years of Interdisciplinary Logistics Research in Bremen, Germany*. Ed. by Michael Freitag, Herbert Kotzab, and Nicole Megow. Cham: Springer International Publishing, pp. 129–153.

Kundu, Shinjini (2021). "AI in medicine must be explainable". In: *Nature Medicine* 27.8, p. 1328. ISSN: 1546170X. DOI: 10.1038/s41591-021-01461-z.

Labov, W. (2006). "Narrative pre-construction." In: *Narrative Inquiry* 16(1).

Laham, Simon M and Yoshihisa Kashima (2013). "Narratives and goals: Narrative structure increases goal priming." In: *Social Psychology* 44.5, p. 303.

Lambrecht, K. (1994). *Information structure and sentence form: topic, focus, and the mental representations of discourse referents*. Cambridge UK: Cambridge University Press.

Lane, Nick (2009). *Life Ascending: The Ten Great Inventions of Evolution*. London: Profile Books.

Larsen, Christian Albrekt (2016). "How three narratives of modernity justify economic inequality". In: *Acta Sociologica* 59.2, pp. 93–111.

Lavine, Jennie S., Ottar N. Bjornsyad, and Rustom Antia (2021). "Immunological Characteristics Govern the Transition of COVID-19 to Endemicity". In: *Science* 371.6530, pp. 741–745. DOI: https://doi.org/10.1126/science.abe6522.

Lebano, Adele and Lynn Jamieson (2020). "Childbearing in Italy and Spain: postponement narratives". In: *Population and Development Review* 46.1, pp. 121–144.

Lefebvre, Georges, Raymond Guyot, and Philippe Sagnac (1951). "La révolution française". In.

Lenat, D. (1995). "CYC: A large-scale investment in knowledge infrastructure". In: *Comm. of the ACMs* 38(11), pp. 33–38.

Levinson, S. (2017). "Speech Acts." In: *The Oxford Handbook of Pragmatics*. Oxford UK: Oxford University Press., pp. 199–216.

Lewis, David, Dennis Rodgers, and Michael Woolcock (2014). *Popular representations of development: Insights from novels, films, television and social media*. Routledge.

Li, W. (2018). *Grounding in Chinese Written Narrative Discourse*. Leiden: Brill.

Lippi, Marco and Paolo Torroni (2016). "Argumentation mining: State of the art and emerging trends". In: *ACM Transactions on Internet Technology (TOIT)* 16.2, pp. 1–25.

Liu, Yushan et al. (2021). "Neural Multi-Hop Reasoning With Logical Rules on Biomedical Knowledge Graphs". In: *European Semantic Web Conference*. Springer, pp. 375–391.

Maller, Cecily, Ralph Horne, and Tony Dalton (2012). "Green renovations: Intersections of daily routines, housing aspirations and narratives of environmental sustainability". In: *Housing, Theory and Society* 29.3, pp. 255–275.

Mat, Paskins and Mary S. Morgan, eds. (2019). *An Anthology of Narrative Science*. Deliverable of the Narrative Science project, funded by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 694732). URL: https://www.narrative-science.org/resources-narrative-science-project.html.

Mateas, Michael and Phoebe Sengers (2003). *Narrative intelligence*. J. Benjamins Pub.

Maurois, André (2017). *Histoire de la France*. La Librairie Vuibert.

McCarthy, J. (1958). "Programs with common sense". In: Symposium on Mechanization of Thought Processes. Teddington, Eng: National Physical Laboratory.

McCarthy, J. and P. Hayes (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence". In: *Machine Intelligence 4*. Ed. by D. Michie and B. Meltzer. Edinburgh University Press, pp. 463–502.

McCloskey, Deirdre (1983). "The rhetoric of economics". In: *Journal of economic literature* 21.2, pp. 481–517.

Meghini, Carlo, Valentina Bartalesi, and Daniele Metilli (2021). "Representing narratives in digital libraries: The narrative ontology". In: *Semantic Web* 18, pp. 1–24.

Meng, Xiao-Li (2021). "Enhancing (publications on) data quality: Deeper data minding and fuller data confession". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Merton, Robert K (1948). "The self-fulfilling prophecy". In: *The antioch review* 8.2, pp. 193–210.

Miller Jr, H Laurence (1962). "On the" Chicago School of Economics"". In: *Journal of Political Economy* 70.1, pp. 64–69.

Minsky, M. (1975). "A framework of representing Knowledge." In: ed. by P. Winston. The Psychology of Computer Vision. New York: McGraw-Hill.

— (1981). "Music, Mind, and Meaning". In: *Computer Music Journal* 5(3).

Mintrom, Michael et al. (2021). "Policy narratives, localisation, and public justification: responses to COVID-19". In: *Journal of European Public Policy* 28.8, pp. 1219–1237.

Mirowski, Philip (1991). *More heat than light: economics as social physics, physics as nature's economics*. Cambridge University Press.

Miskimmon, Alister, Ben O'loughlin, and Laura Roselle (2014). *Strategic narratives: Communication power and the new world order*. Routledge.

Miskimmon, Alister, Ben O'Loughlin, and Laura Roselle (2015). "Strategic narratives: A response". In: *Critical Studies on Security* 3.3, pp. 341–344.

Mitchell, M. (2019). *Artificial Intelligence. A guide for thinking humans*. New York: Farrar, Strauss and Giroux.

Mohseni, S., N. Zarei, and E. Ragan (2021). "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems". In: *ACM Trans. Interact. Intell. Syst.* 1(1).

Molnar, Christoph et al. (2020). "Pitfalls to avoid when interpreting machine learning models". In: *arXiv preprint arXiv:2007.04131*.

Monaco, J. (2000). *How to read a film*. Oxford: Oxford University Press.

Montabone, S. and A. Soto (2010). "Human Detection Using a Mobile Platform and Novel Features Derived From a Visual Saliency Mechanism." In: *Image and Vision Computing* 28(3), 391–402.

Moore, J. and W. Swartout. (1988). *Explanation in Expert Systems: A survey*. Marina del Rey, Cal.

Morgan, Mary S (2001). "Models, stories and the economic world". In: *Journal of Economic Methodology* 8.3, pp. 361–384.

Morgan, Mary S. and Matthew Norton Wise (2017). "Narrative Science and Narrative Knowing. Introduction to Special Issue on Narrative Science". In: *Studies in History and Philosophy of Science Part A* 62, pp. 1–5. DOI: https://doi.org/10.1016/j.shpsa.2017.03.005.

Morrell, Kevin and Chanaka Jayawardhena (2010). "Fair trade, ethical decision making and the narrative of gender difference". In: *Business Ethics: A European Review* 19.4, pp. 393–407.

Murray, Janet (1998). *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge MA: MIT Press.

Nakai, Asako (2020). "Narratives of Inequality: Postcolonial Literary Economics." In: *The Journal of New Zealand Studies* NS30.

Nowak, A., P. Lukowicz, and P. Horodecki (2018). "Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance? or the Biggest Threat". In: *IEEE Technology and Society Magazine* 37(4), pp. 26–34.

Nyman, Rickard, Sujit Kapadia, and David Tuckett (2021). "News and narratives in financial systems: exploiting big data for systemic risk assessment". In: *Journal of Economic Dynamics and Control* 127, p. 104119.

Palgrave, Robert Harry Inglis (1894). *Dictionary of Political Economy*. en. Macmillan and Company.

Palmer, Robert Roswell and David Armitage (2014). *The Age of the Democratic Revolution: A Political History of Europe and America, 1760–1800*. Updated Edition with a New Foreword. Princeton NJ: Princeton University Press.

Palvia, Prashant, Naveed Baqir, and Hamid Nemati (2018). "ICT for socio-economic development: A citizens' perspective". In: *Information & Management* 55.2, pp. 160–176.

Panichello, M. and T. Buschman (2021). "Shared mechanisms underlie the control of working memory and attention." In: *Nature* 592(7855), pp. 601–605.

Panofsky, E. (1939,1972). *Studies in Iconology. Humanistic themes in the art of the Renaissance*. Oxford: Oxford University Press.

Pavey, E. (2010). *The Structure of Language: An Introduction to Grammatical Analysis*. Cambridge, UK: Cambridge University Press.

Pearl, J. and D. Mackenzie (2019). *The book of why. The new science of cause and effect*. London: Penguin Books.

Percha, Bethany and Russ B Altman (2018). "A global network of biomedical relationships derived from text". In: *Bioinformatics* 34.15, pp. 2614–2624.

Phelan, James (1996). *Narrative as rhetoric: Technique, audiences, ethics, ideology*. Ohio State University Press.

Pickering, S., J. McCulloch, and D. Wright-Neville (2008). "Counter-terrorism policing: Towards social cohesion." In: *Crime, Law and Social Change* 50(1-2).

Pikler, Andrew G (1954). "Utility theories in field physics and mathematical economics (I)". In: *The British Journal for the philosophy of Science* 17, pp. 47–58.

— (1955). "Utility theories in field physics and mathematical economics (II)". In: *The British Journal for the Philosophy of Science* 20, pp. 303–318.

Polletta, F. and M. Kai Ho (2021). "Frames and their consequences." In: *The Oxford handbook of contextual political analysis*. 5, pp. 1–15.

Poltorak, Mike et al. (2005). "'MMR talk'and vaccination choices: An ethnographic study in Brighton". In: *Social Science & Medicine* 61.3, pp. 709–719.

Porzel, R. (2021). "On Formalizing Narratives". In: *CAOS 2021: 5th workshop on Cognition and Ontologies*. CEUR Workshop Proceedings. Bolzano.

Preece, Chloe and Finola Kerrigan (2015). "Multi-stakeholder brand narratives: An analysis of the construction of artistic brands". In: *Journal of Marketing Management* 31.11-12, pp. 1207–1230.

Pretus, C., N. Hamid, H. Sheikh, J. Ginges, et al. (2018). "Neural and behavioral correlates of sacred values and vulnerability to violent extremism." In: *Frontiers in Psychology* 9(DEC), p. 2462.

Pretus, C., N. Hamid, H. Sheikh, A. Gomez, et al. (2019). "Ventromedial and dorsolateral prefrontal interactions underlie will to fight and die for a cause." In: *Social Cognitive and Affective Neuroscience* 14(6), pp. 569–77.

Quillian, R. (1968). "Semantic Memory". In: *Semantic Information Processing*. Ed. by M. Minsky. Cambridge Ma: The MIT Press. Chap. 4, pp. 227–270.

Reddy, R. et al. (1974). "The HEARSAY Speech Understanding System." In: *Journal of the Acoustical Society of America*. 55, p. 409.

Redmon, Joseph et al. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. arXiv: 1506.02640 [cs.CV].

Riedl, Mark O. (2016). "Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence". In: *CHI 2016 Workshop on Human-Centered Machine Learning* abs/1602.06484. arXiv: 1602.06484.

— (2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34, 49.

Rodden, John (2008). "How do stories convince us? Notes towards a rhetoric of narrative". In: *College Literature*, pp. 148–173.

Rojas, Diego, Carlos A Vegh, and Guillermo Vuletin (2020). *The macroeconomic effects of macroprudential policy: evidence from a narrative approach*. Tech. rep. National Bureau of Economic Research.

Romer, Christina D and David H Romer (2004). "A new measure of monetary shocks: Derivation and implications". In: *American Economic Review* 94.4, pp. 1055–1084.

— (2010). "The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks". In: *American Economic Review* 100.3, pp. 763–801.

Roos, Michael and Matthias Reccius (2021). "Narratives in economics". In: *arXiv preprint arXiv:2109.02331*.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

Safari, Aswo and Peter Thilenius (2013). "Alleviating uncertainty through trust: A narrative approach to consumers' foreign online purchasing behaviour". In: *Journal of customer behaviour* 12.2-3, pp. 211–226.

Saltelli, Andrea et al. (2020). "The technique is never neutral. How methodological choices condition the generation of narratives for sustainability". In: *Environmental Science & Policy* 106, pp. 87–98.

Samuel, A. (1959). "Some studies in machine learning using the game of checkers". In: *IBM Journal of Research and Development* 3(3), pp. 210–229.

Sandra, D., J. Verschueren, and J-O Ostman (eds.) (2009). *Cognition and Pragmatics*. Amsterdam: John Benjamins Pub.

Santagiustina, Carlo R.M.A. (2018). "Talking about uncertainty". In.

— (2022). "From Narrative Economics to Economics' Narratives". In: *MUHAI Deliverable 1.1: Foundations for Incorporating Meaning and Understanding in Human-Centric AI*. Bremen: The MUHAI Consortium.

Schacter, Daniel L, Donna Rose Addis, and Randy L Buckner (2007). "Remembering the past to imagine the future: the prospective brain". In: *Nature reviews neuroscience* 8.9, pp. 657–661.

Schank, R. (1990). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge Eng: Cambridge University Press.

Schank, R. and Abelson (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: L. Erlbaum.

Schiller, R. (2019). *Narrative Economics. How Stories Go Viral and Drive Major Economic Events*. Princeton, NJ: Princeton University Press.

Scholz, Oliver R (2018). "Symptoms of expertise: knowledge, understanding and other cognitive goods". In: *Topoi* 37.1, pp. 29–37.

Sedláček, Tomáš (2014). "Economics of good and evil". In: *Bonds*. Wilhelm Fink, pp. 331–342.

Shane, J. (2021). *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place*. New York: Little Brown and Company.

Sheikh, H. et al. (2012). "Religion, group threat and sacred values." In: *Judgment and Decision Making* 7(2), pp. 110–118.

Shiller, Robert J (2020). *Narrative economics*. Princeton University Press.

— (2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004. DOI: 10.1257/aer.107.4.967. URL: https://www.aeaweb.org/articles?id=10.1257/aer.107.4.967.

Simon, H. (1969). *The Sciences of the Artificial*. Cambridge Ma: The MIT Press.

Singer, Tania (2015). "From homo economicus towards a computational model of caring economics: How motivation determines decision making and cooperation". In: *International Convention for Psychological Science (ICPS)*.

Slatin, Sonia (1979). "Opera and Revolution: La Muette de Portici and the Belgian Revolution of 1830 Revisited". In: *Journal of Musicological Research* 3, pp. 45–62. DOI: https://doi.org/10.1080/01411897908574506.

Sools, A., T. Tromp, and J. Mooren (2015). "Mapping letters from the future: Exploring narrative processes of imagining the future". In: *Journal of Health Psychology* 20(3), pp. 350–364.

Sosa, Daniel N et al. (2019). "A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*. World Scientific, pp. 463–474.

Soulet, Arnaud et al. (2018). "Representativeness of knowledge bases with the generalized Benford's law". In: *International Semantic Web Conference*. Springer, pp. 374–390.

Sperber, D. and D. Wilson (1969). *Relevance: Communication and Cognition*. London: Wiley-Blackwell.

Steels, L. (1991). "Emergent frame recognition and its use in artificial creatures". In: *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*. IJCAI, pp. 1219–1224.

— (2004). "Constructivist Development of Grounded Construction Grammars". In: *Proceedings Annual Meeting of Association for Computational Linguistics Conference*. Association for Computational Linguistics, pp. 9–16.

— (2012a). *Computational Issues in Fluid Construction Grammar*. Vol. 7249. Lecture Notes in Computer Science. Berlin: Springer Verlag.

— (2012b). *Experiments in Cultural Language Evolution*. Amsterdam: John Benjamins, Pub.

Steels, L. (2016). "Agent-based models for the emergence and evolution of grammar." In: *Phil. Trans. R. Soc. B.* 371: 20150447.

— (2020). "Personal Dynamic Memories are Necessary to Deal with Meaning and Understanding in Human-Centric AI." In: *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) Co-located with 24th European Conference on Artificial Intelligence*. Ed. by L. Serafini Saffiotti A and P. Lukowicz. Vol. Vol-2659. CEUR Workshop Proceedings.

— (2021). "From Audio Signals to Musical meaning." In: *Handbook of Artificial Intelligence for Music. Foundations, Advanced Approaches, and Developments for Creativity*. Ed. by E. Miranda. Berlin: Springer-Verlag, pp. v–xviii.

— (2022). "Conceptual Foundations of Human-Centric AI." In: *Advanced course on Human-Centered AI. ACAI 2021*. Ed. by M. Chetouani et al. Springer Lecture Notes in Artificial Intelligence (LNAI) Post-Proceedings Volume, Tutorial Lecture Series. Berlin: Springer Verlag.

Steels, L. and R. Brooks (1994). *The 'artificial life' route to 'artificial intelligence'. Building Situated Embodied Agents*. New Haven: Lawrence Erlbaum Ass.

Steels, L. and M. Hild (2012). *Language grounding in robots*. New York: Springer Verlag.

Steels, L. and B. Wahle (2019). "Perceiving the Focal Point of a Painting with AI. Case Studies on works of Luc Tuymans." In: *12th International Conference on Agents and Artificial Intelligence*. Setubal, Portugal: Scite Press.

Steels, Luc (1998). "The origins of ontologies and communication conventions in multi-agent systems". In: *Autonomous Agents and Multi-Agent Systems* 1, pp. 169–194.

— (2011). *Design patterns in fluid construction grammar*. Vol. 11. John Benjamins Publishing.

— (2022). "Conceptual Foundations of Human-Centric AI". In: *MUHAI Deliverable 1.1: Foundations for Incorporating Meaning and Understanding in Human-Centric AI*. Bremen: The MUHAI Consortium.

Steinhardt, Joseph and Michael A Shapiro (2015). "Framing effects in narrative and non-narrative risk messages". In: *Risk Analysis* 35.8, pp. 1423–1436.

Stravinsky, I. (1935). *An autobiography*. New York: Simon and Schuster.

Svetlova, Ekaterina (2021). "AI meets narratives: the state and future of research on expectation formation in economics and sociology". In: *Socio-Economic Review*.

Szegedy, C. et al. (2014). "Intriguing properties of neural networks." In: *International Conference on Learning Representations*. Springer, pp. 159–166. arXiv: 1312.6199.

Teeter, Preston and Jörgen Sandberg (2017). "Cracking the enigma of asset bubbles with narratives". In: *Strategic Organization* 15.1, pp. 91–99.

Tegmark, M. (2017). *Life 3.0: being human in the age of artificial intelligence*. New York: Knopff.

The Florentine Catasto of 1427 (n.d.). Datastory at https://stories.datalegend.net/catasto/, last retrieved on February 17, 2022.

"The myth of generalisability in clinical research and machine learning in health care" (2020). In: *The Lancet Digital Health* 2.9, e489–e492. ISSN: 25897500. DOI: 10.1016/

S2589-7500(20)30186-2. URL: http://dx.doi.org/10.1016/S2589-7500(20)30186-2.

Thorndyke, Perry W (1977). "Cognitive structures in comprehension and memory of narrative discourse". In: *Cognitive psychology* 9.1, pp. 77–110.

Tiddi, I., M. D'Aquin, and E. Motta (2014). "Walking linked data: A graph traversal approach to explain clusters." In: *Proceedings of the 5th International Workshop on Consuming Linked Data (COLD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014)*. Ed. by J. Glasgow, N. Narayanan, and B. Chandrasekaran. CEUR Workshop proceedings.

Tiddi, Ilaria, Mathieu D'Aquin, and Enrico Motta (2014). "Walking linked data: A graph traversal approach to explain clusters". In: *CEUR Workshop Proceedings*.

Tiddi, Ilaria, Mathieu d'Aquin, and Enrico Motta (2014). "Quantifying the bias in data links". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 531–546.

Tilly, Sonja, Markus Ebner, and Giacomo Livan (2021). "Macroeconomic forecasting through news, emotions and narrative". In: *Expert Systems with Applications* 175, p. 114760.

Titumir, Rashed Al Mahmud (2021). *Numbers and Narratives in Bangladesh's Economic Development*. Springer Nature.

Tomlin, R. and A. Myachykov (2015). "Attention and salience." In: *Handbook of Cognitive Linguistics*. Ed. by E. Dabrowska and D. Divjak. Berlin: De Gruyter., pp. 31–52.

Tuckett, David and David Tuckett (2011). *Minding the markets: An emotional finance view of financial instability*. Springer.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under uncertainty: Heuristics and biases". In: *science* 185.4157, pp. 1124–1131.

Van Der Leeuw, Sander (2020). "The role of narratives in human-environmental relations: an essay on elaborating win-win solutions to climate change and sustainability." In: *Climatic Change* 160.4.

Van Eecke, P., J. Nevens, and K. Beuls (2022). "Neural Heuristics for Scaling Constructional Language Processing." In: *Submitted*.

Van Harmelen, Frank, Vladimir Lifschitz, and Bruce Porter (2008). *Handbook of knowledge representation*. Elsevier.

Van Noort, Carolijn (2017). "Study of strategic narratives: The case of BRICS". In: *Politics and Governance* 5.3, pp. 121–129.

Vandelanotte, L. (2009). "Introduction." In: *Key Notions for Pragmatics*. Ed. by J. Verschueren. Amsterdam: John Benjamins Pub., pp. 1–27.

— (2017). "Attention and salience." In: *The Cambridge Handbook of Cognitive Linguistics*. Ed. by B. Dancygier. Cambridge, UK: Cambridge University Press., pp. 157–171.

Vignoli, Daniele et al. (2020). "A reflection on economic uncertainty and fertility in Europe: The Narrative Framework". In: *Genus* 76.1, pp. 1–27.

Vilarroya, O. (2019). *Somos lo que nos contamos. Cómo los relatos construyen el mundo en que vivimos*. Barcelona: Editorial Planeta.

Von der Leyen, U. and et al. (2020). "White paper on Artificial Intelligence". In: *EU Commission reports*.

Walker, Donald A (1991). "Economics as social physics". In: *The Economic Journal* 101.406, pp. 615–631.

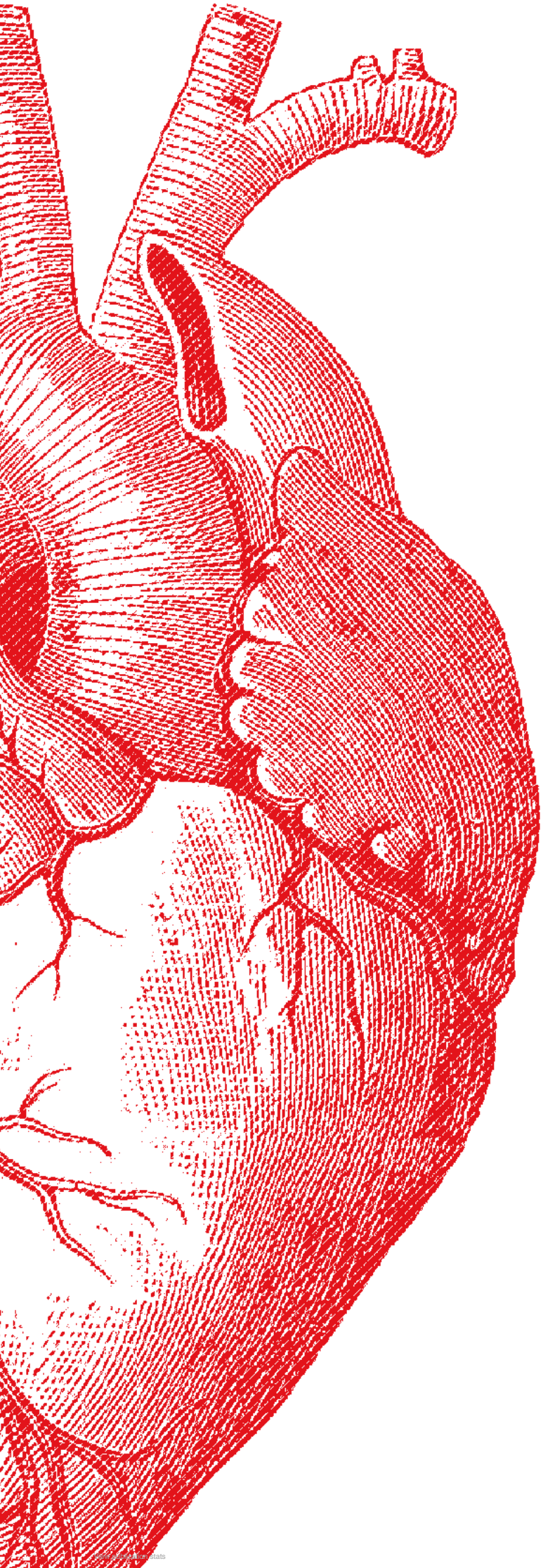
White, Hayden (1981). "The narrativization of real events". In: *Critical Inquiry* 7.4, pp. 793–798.

Willems, R., S. Nastase, and B. Milivojevic (2020). "Narratives for Neuroscience." In: *Trends in Neurosciences* 44(5), pp. 271–273.

Winograd, T. (1976). *Understanding Natural Language*. New York: Academic Press.

Winston, P. (2011). "The Strong Story Hypothesis and the Directed Perception Hypothesis". In: 2011 AAAI Fall symposium. Menlo Park Ca: AAAI Press.

Zaloom, Caitlin (2003). "Ambiguous numbers: Trading technologies and interpretation in financial markets". In: *American ethnologist* 30.2, pp. 258–272.



Current AI rests almost exclusively on statistically acquired pattern recognition and pattern completion. But there is now a consensus that human-centric AI, which requires better explanation, human influence, more robustness, and explicit ethical foundations, needs to go beyond pattern recognition to incorporate meaning and understanding.

Because meaning and understanding are rather vague and overloaded notions there is no obvious research path to achieve it. This volume explores how the notion of understanding is being discussed and treated in other human-centred research fields, more specifically in social brain science, social psychology, linguistics, semiotics, economics, social history and medicine, in order to gather ideas for building the next generation of human-centric AI and understand what technology gaps need to be plugged.

This project has received funding from the European Pathfinder Project under Grant Agreement N° 951846

ISBN 979-12-810-8700-2