

Generalization by People and Machines

Barbara Hammer^{*1}, Filip Ilievski^{*2}, Sascha Saralajew^{*3}, and Frank van Harmelen^{*4}

1 Universität Bielefeld, DE, bhammer@techfak.uni-bielefeld.de

2 VU Amsterdam, NL, f.ilievski@vu.nl

3 NEC Laboratories Europe – Heidelberg, DE, sascha.saralajew@neclab.eu

4 VU Amsterdam, NL, frank.van.harmelen@vu.nl

Abstract

Today's AI systems are powerful to the extent that they have largely entered the mainstream and divided the world between those who believe AI will solve all our problems and those who fear that AI will be destructive for humanity. Meanwhile, trusting AI is very difficult given its lack of robustness to novel situations, consistency of its outputs, and interpretability of its reasoning process. Building trustworthy AI requires a paradigm shift from the current oversimplified practice of crafting accuracy-driven models to a human-centric design that can enhance human ability on manageable tasks, or enable humans and AIs to solve complex tasks together that are difficult for either separately. At the core of this problem is the unrivaled human generalization and abstraction ability. While today's AI is able to provide a response to any input, its ability to transfer knowledge to novel situations is still limited by oversimplification practices, as manifested by tasks that involve pragmatics, agent goals, and understanding of narrative structures. As there are currently no venues that allow cross-disciplinary research on the topic of reliable AI generalization, this discrepancy is problematic and requires dedicated efforts to bring in one place generalization experts from different fields within AI, but also with Cognitive Science. This Dagstuhl Seminar thus provided a unique opportunity for discussing the discrepancy between human and AI generalization mechanisms and crafting a vision on how to align the two streams in a compelling and promising way that combines the strengths of both. To ensure an effective seminar, we brought together cross-disciplinary perspectives across computer and cognitive science fields. Our participants included experts in Interpretable Machine Learning, Neuro-Symbolic Reasoning, Explainable AI, Commonsense Reasoning, Case-based Reasoning, Analogy, Cognitive Science, and Human-AI Teaming. Specifically, the seminar participants focused on the following questions: How can cognitive mechanisms in people be used to inspire generalization in AI? What Machine Learning methods hold the promise to enable such reasoning mechanisms? What is the role of data and knowledge engineering for AI and human generalization? How can we design and model human-AI teams that can benefit from their complementary generalization capabilities? How can we evaluate generalization in humans and AI in a satisfactory manner?

Seminar May 5–8, 2024 – <https://www.dagstuhl.de/24192>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Cognitive science

Keywords and phrases Abstraction, Cognitive Science, Generalization, Human-AI Teaming, Interpretable Machine Learning, Neuro-Symbolic AI

Digital Object Identifier 10.4230/DagRep.14.5.1

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Generalization by People and Machines, *Dagstuhl Reports*, Vol. 14, Issue 5, pp. 1–11
Editors: Barbara Hammer, Filip Ilievski, Sascha Saralajew, and Frank van Harmelen



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Filip Ilievski (VU Amsterdam, NL)

Sascha Saralajew (NEC Laboratories Europe – Heidelberg, DE)

License  Creative Commons BY 4.0 International license
© Filip Ilievski and Sascha Saralajew

The Dagstuhl Seminar consisted of

1. lightning talks, where each participant had 2min for a short introduction and the presentation of a motivating (funny) example of generalization,
2. perspective pitches, where invited researchers from different domains gave a short talk about generalization in their domain (15min talk and 15min discussion),
3. daily discussion breakout sessions, where researchers organized in groups to discuss aspects of generalization and to work on the joint perspectives paper, and
4. plenary sessions, where we discussed the progress and results of the different breakout groups and organizing question around the seminar.

Day 1 featured an introductory session by the organizers and the lightning talks. During day 1, there were two perspective pitches on generalization from the angle of analogy (by *Ken Forbus*) and knowledge representation in symbolic AI (by *Luciano Serafini*). In the afternoon of day 1, the participants discussed generalization in four working groups:

1. types of generalization,
2. methods of generalization,
3. evaluation of generalization, and
4. human-AI teaming.

All teams were comprised of participants with diverse background and interests. The formation of these four groups was informed by a poll on possible topics that was sent to the participants before the seminar, filled by nearly all participants. At the end of day 1, each group provided updates centered around three questions:

1. How is it done today?
2. How well are we doing?
3. What are open challenges and important future directions?

Day 1 ended up with a plenary session during which each of the groups reported on their initial ideas, and received feedback from the other participants.

Day 2 featured four perspective pitches, highlighting the angles of statistical physics (by *Michael Biehl*), cognitive science (by *Ute Schmid*), computational linguistics (by *Vered Shwartz*), and computer vision (by *Wael AbdAlmageed*). In the afternoon, the participants split into the same four working groups as in day 1, with an instruction to organize the list of considerations from day 1. A key goal was to narrow down the scope of each working group and to identify important points to focus on. Day 2 ended with a debrief by the breakout sessions, during which common aspects emerged in different groups.

To facilitate a fruitful end of the seminar, the organizers came up with a set of 4 pillars that each of the groups was supposed to organize their content around, during day 3. These included: theory, context, representation, and foundational models. On day 3, each group provided an attempt to organize their content into these four pillars to the extent possible. Day 3 (half a day) ended with a discussion on the next steps, with a specific goal of writing a joint agenda-setting paper with all participants, targeted at a prestigious venue.

In the meantime, the organizers and Prof. Ute Schmid formed an editorial team that has been leading the process of writing the perspectives paper, and the participants provided two versions of write-up from their group: a long version and a short version. The short versions are limited to 2-3 pages and 20-25 citations, to conform jointly with the restrictions of journals like Nature Machine Intelligence. At the time of writing, the editorial team is busy with preparing this submission, with another round of feedback and collaboration scheduled with the participants in August.

2 Table of Contents

Executive Summary
Filip Ilievski and Sascha Saralajew 2

Overview of Talks

Generalization from the perspective of computer vision
Wael Abd-Almageed 5

Generalization from the perspective of the statistical physics of learning
Michael Biehl 5

Generalization in People and Machines: An Analogy/Cognitive Science Perspective
Kenneth D. Forbus 6

Generalization and Abstraction in Cognitive Science
Ute Schmid 6

What do knowledge representation people think when they hear “generalisation”
Luciano Serafini and Frank van Harmelen 6

Generalization from the perspective of language
Vered Shwartz 7

Working groups

Methods of Generalization
Barbara Hammer, Xin Luna Dong, Giuseppe Marra, Axel-Cyrille Ngonga Ngomo, Gabriella Pasi, Dafna Shahaf, and Frank van Harmelen 8

Human-AI Teaming and Generalization
Pascal Hitzler, Alessandro Oltramari, Zeynep G. Saribatur, Ute Schmid, John Shawe-Taylor, Gabriella Skitalinska, Clemens Stachl, Piek Vossen, and Michael R. Waldmann 8

Evaluation of Generalization
Filip Ilievski, Kiril Gashteovski, Pasquale Minervini, Martin Mundt, Sascha Saralajew, and Vered Shwartz 9

Types of Generalization
Benjamin Paaßen, Wael Abd-Almageed, Michael Biehl, Marianna Marcella Bolognesi, Kenneth D. Forbus, Luciano Serafini, Gido van de Ven, and Thomas Villmann 10

Participants 11

3 Overview of Talks

3.1 Generalization from the perspective of computer vision

Wael Abd-Almageed (Clemson University, US)

License  Creative Commons BY 4.0 International license
© Wael Abd-Almageed

Artificial Intelligence (AI) has been experiencing significant advances in the last 10 years, including image and video understanding and natural language dialog systems. AI promises to disrupt a wide range of applications and industries from self-driving cars and intelligent transportation to drug discovery and healthcare. However, existing AI technology suffer from major limitations in terms of generalization to real-world scenarios and real-world data. In the first part of this talk, I will be discussing several limitations of computer vision systems, as one important modality of AI systems. For example, the performance of computer vision systems trained to classify, detect and/or segment a set of object classes degrades rapidly when these systems are deployed in new environments where the statistical distribution and/or characteristics of the data is different than training data. Meanwhile, computer vision systems with continual learning capabilities struggle to differentiate between outliers of known classes (e. g., unusual fish or bird) and samples from completely new classes (e. g., new biometric face spoofing attack) that should be incorporated into the AI system. Further, continual learning system often suffer from catastrophic forgetting, when learning new classes and/or adapting to new data distributions leads to performance degradation on already learned classes/distributions. In the second part of the talk, I will discuss a hybrid NeuroSymbolic artificial intelligence architecture that mitigates the limitations of existing AI systems and leads to better generalization and reasoning capabilities, when AI systems are deployed in new real-world environments

3.2 Generalization from the perspective of the statistical physics of learning


Michael Biehl (University of Groningen, NL)

License  Creative Commons BY 4.0 International license
© Michael Biehl

In this presentation, the term generalization refers to the ability of adaptive systems, for instance, neural networks, to apply a rule that is learned from training examples to novel, unseen data in the working phase. The statistical physics approach to learning theory is outlined very briefly. It complements other theoretical frameworks and has re-gained significant interest due to the growing popularity of neural networks and machine learning in general. The computation of typical learning curves in so-called student teacher model scenarios is exemplified in terms of training layered networks by stochastic optimization of an objective function. Assuming training from randomized data sets, the average generalization ability is computed as a function of the training set size, for instance, the number of available examples. As an important example result, the existence of phase transitions in batch training is discussed: here the generalization ability improves suddenly at a critical data set size. Similarly, the analysis of the training dynamics of stochastic gradient descent reveals the existence of plateau states which can dominate the training process. They are left by means of rapid changes of the generalization ability with time and can lead to cascade-like learning curves.

3.3 Generalization in People and Machines: An Analogy/Cognitive Science Perspective


Kenneth D. Forbus (Northwestern University – Evanston, US)

License  Creative Commons BY 4.0 International license
© Kenneth D. Forbus

How does the human ability to generalize work? This talk examined two sources of this capability. The first are qualitative representations, which provide abstract causal and spatial models that are easier to learn than detailed quantitative models. The second is analogy, where the process of analogical matching provides a means of constructing generalizations by identifying what is common across a set of examples. The talk outlined Gentner’s structure-mapping theory, the analogy stack consisting of cognitive models of matching, retrieval, and generalization, and how these have been used in a variety of cognitive simulations and performance-oriented AI systems. The Continuum of Knowledge Hypothesis was discussed, which proposes that knowledge starts out concrete and is incrementally and partially abstracted in stages. Finally, a set of open questions was discussed.

3.4 Generalization and Abstraction in Cognitive Science


Ute Schmid (Universität Bamberg, DE)

License  Creative Commons BY 4.0 International license
© Ute Schmid

Generalization is defined as transfer of what has been learned in one context to a new one which is similar. Representation and how similarity is assessed are crucial for generalization. Generalization can involve abstraction of general characteristics (deleting irrelevant and constructing more general features) for a collection of entities. In the talk I give an introduction to classic theoretical approaches and empirical findings from cognitive science with a focus on concept learning. Open questions, from my perspective, are: (1) The relationship between generalisation and representation: Where does structure come from? What is the human inductive bias which leads to useful generalizations? (2) What is the relation between implicit and explicit learning?

3.5 What do knowledge representation people think when they hear “generalisation”

Luciano Serafini (Bruno Kessler Foundation – Trento, IT) and Frank van Harmelen (VU Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Luciano Serafini and Frank van Harmelen

Generalization in KRR is usually defined with respect to a logical framework where background knowledge, also known as inductive biases, is expressed through sentences of a logical language. The adopted logical framework provides an **inference mechanism** to check logical consequence and a **background theory (set of formulas)** to explicitly state assumptions (inductive biases). Within this framework, four main processes of generalization can be

formally defined. The first process is **Predicate Invention**: this involves identifying a set of elements in the domain of interest and defining the necessary and sufficient conditions for membership in this set. The second process of generalization occurs during **Clustering**: given a set of individuals S with associated properties, the goal is to find a partition S_1, \dots, S_k of S based on the similarity of their properties. Subsequently, extend the language with new symbols for each cluster in the partition. Another process of generalization is **Subsumption**: starting from classes C_1, \dots, C_k , introduce a superclass S that **subsumes** each C_i , meaning that every instance of C_i is an instance of S , and optionally, every instance of S is an instance of some C_i . A further generalization method is called **Rule Mining**: given a set F of **ground facts** about a subset of individuals S , find a set of lifted rules that hold for a larger set of individuals $S' \supset S$. **Building Analogies** is another generalization process found in the KR literature. In this case, given a base domain B and a target domain T , find a mapping α between the objects of B and T that preserves relational structure. Finally, the operation of **extending a formal theory** is also a generalization process where a theory T is expanded by adding new symbols to the language of T , providing a new set of axioms T' that relate the new symbols to the existing ones, and then identifying a condition C such that: $T \models \phi$ if and only if $C, T' \models \phi$.

3.6 Generalization from the perspective of language

Vered Shwartz (University of British Columbia – Vancouver, CA)

License © Creative Commons BY 4.0 International license
© Vered Shwartz

Out-of-distribution generalization in natural language processing is the ability of models to solve examples from a different distribution of the training data, based on prior knowledge and similarity to training examples. This includes robustness to prediction when introducing superficial changes to the input; and updating the prediction when introducing semantic changes to the input. Lack of generalization makes models brittle and unsafe to deploy for real-world applications. Current evaluation methods for generalization include cross-dataset evaluations and adversarial examples. In terms of making models more generalizable, there are several model enhancements such as partial model updates, neuro-symbolic and compositional models, and training on fewer examples (such as few-shot learning). From the data perspective, training on more data or specifically on adversarial examples can make models more robust. LLMs are exceptionally general and versatile, given their training on vast amounts of raw text. They are to some extent able to generalize to new concepts and ideas. However, they still over-rely on similar training examples, and are brittle when these examples are manipulated. They are still not robust to changes in the prompt phrasing. There is no evidence that they are capable of causal reasoning. Finally, testing generalization in LLMs is tricky without access to the training data.

4 Working groups

4.1 Methods of Generalization

Barbara Hammer (Universität Bielefeld, DE), Xin Luna Dong (Meta Reality Labs – Bellevue, US), Giuseppe Marra (KU Leuven, BE), Axel-Cyrille Ngonga Ngomo (Universität Paderborn, DE), Gabriella Pasi (University of Milan, IT), Dafna Shahaf (The Hebrew University of Jerusalem, IL), and Frank van Harmelen (VU Amsterdam, NL)

License © Creative Commons BY 4.0 International license
 © Barbara Hammer, Xin Luna Dong, Giuseppe Marra, Axel-Cyrille Ngonga Ngomo, Gabriella Pasi, Dafna Shahaf, and Frank van Harmelen

The group on methods for generalization identified three families of methods:

1. symbolic, for instance, predicate invention, semantic clustering, subsumption, rule mining,
2. statistical, for instance, machine learning methods, generative AI, representation learning, and
3. combinations of the two, for instance, neuro-symbolic methods, embedding-based methods.

Generalizations can be learned directly from data, or they can be obtained from a combination of data and knowledge. Key considerations about methods include

- provable properties of generalizations including worst-case guarantees,
- context sensitivity of generalization,
- methods for explainability,
- compositionality,
- quantifying the trade-off between compression, memorization and forgetting,
- evolving generalizations over time, and
- choice of appropriate representations.

This group was coordinated by Frank van Harmelen and Barbara Hammer.

4.2 Human-AI Teaming and Generalization

Pascal Hitzler (Kansas State University – Manhattan, US), Alessandro Oltramari (Carnegie Bosch Institute – Pittsburgh, US), Zeynep G. Saribatur (TU Wien, AT), Ute Schmid (Universität Bamberg, DE), John Shawe-Taylor (University College London, GB), Gabriella Skitalinska (Leibniz Universität Hannover, DE), Clemens Stachl (Universität St. Gallen, CH), Piek Vossen (VU Amsterdam, NL), and Michael R. Waldmann (Universität Göttingen, DE)

License © Creative Commons BY 4.0 International license
 © Pascal Hitzler, Alessandro Oltramari, Zeynep G. Saribatur, Ute Schmid, John Shawe-Taylor, Gabriella Skitalinska, Clemens Stachl, Piek Vossen, and Michael R. Waldmann

Machine-learning based systems and humans both are capable of generalizing from examples. However, generalization capabilities appear to differ significantly, with complementary strengths and weaknesses. For example, humans are generally good at commonsense reasoning, using structured knowledge, and handling out-of-distribution data. Machine learning excels at objectivity (at least based on the data given), at scale, and at high complexity. This complementarity gives opportunities for human-machine teaming, with each side addressing the limitations of the other. For example, some generalization capabilities of LLMs, like the quick production of rhetorically polished texts on any topic, are beyond that of most humans. Yet, they make generalization errors (called “hallucinations”) like

the replacing of specific facts with non-factual information; an error easily caught by a knowledgeable human. But such human-machine teaming breaks down if the human is not a topic expert.

The need for teaming arises naturally in complex application scenarios, for instance, automotive driver assistance or complex decision making. For these, it is of central importance that the human can assess machine responses, for example, has access to the rationales (called “explanations”) on the basis of which the machine responded. Future XAI research must prioritize understanding human cognition because effective human-AI collaboration requires explanations that bridge the explanatory gap between human reasoning and AI’s internal workings.

A critical challenge lies in reconciling fundamentally different reasoning paradigms: human causal models versus AI’s deep learning associations. Can these approaches be unified into a common explanatory language? Furthermore, fostering successful human-AI teams necessitates AI’s ability to learn and potentially retain feedback indefinitely. Robust feedback mechanisms are crucial for AI to understand effective communication and align with human cognition, fostering seamless collaboration. Future research should also prioritize the investigation of human generalization and abstraction processes and contrast those with AI-based approaches. Interdisciplinary collaboration between computer and social sciences will be essential to integrate this understanding into AI design, not only enhancing explainability but also mitigating biases in machine learning generalization. This group was coordinated by Pascal Hitzler, with help from John Shawe-Taylor.

4.3 Evaluation of Generalization

Filip Ilievski (VU Amsterdam, NL), Kiril Gashteovski (NEC Laboratories Europe – Heidelberg, DE), Pasquale Minervini (University of Edinburgh, GB), Martin Mundt (TU Darmstadt, DE), Sascha Saralajew (NEC Laboratories Europe – Heidelberg, DE), and Vered Shwartz (University of British Columbia – Vancouver, CA)

License © Creative Commons BY 4.0 International license
© Filip Ilievski, Kiril Gashteovski, Pasquale Minervini, Martin Mundt, Sascha Saralajew, and Vered Shwartz

The generalization group identified certain challenges with evaluation of generalization, such as:

- inadequate data splitting practices,
- hard to define the bounds of generalization,
- no comprehensive way to evaluate the total phenomena,
- limited metrics,
- consolidation challenges of discriminative and generative evaluation,
- measuring tradeoffs between predictive power and efficiency,
- evaluating long-tail phenomena, and
- selection of the right granularity for generalization.

Emerging practices for evaluation include

- cross-benchmark evaluations,
- testing robustness to perturbations,
- evaluations of over- and under-generalization,
- evaluation with multiple metrics, and
- factoring out memorization.

Many important questions were identified as relevant future works, including

- the design of checklists,
- evaluations of different levels of similarity,
- clear definition of bounds of generalization, and
- quantification of variations in performance.

This group was coordinated by Filip Ilievski and Sascha Saralajew.

4.4 Types of Generalization

Benjamin Paaßen (Universität Bielefeld, DE), Wael Abd-Almageed (Clemson University, US), Michael Biehl (University of Groningen, NL), Marianna Marcella Bolognesi (University of Bologna, IT), Kenneth D. Forbus (Northwestern University – Evanston, US), Luciano Serafini (Bruno Kessler Foundation – Trento, IT), Gido van de Ven (KU Leuven, BE), and Thomas Villmann (Hochschule Mittweida, DE)

License © Creative Commons BY 4.0 International license

© Benjamin Paaßen, Wael Abd-Almageed, Michael Biehl, Marianna Marcella Bolognesi, Kenneth D. Forbus, Luciano Serafini, Gido van de Ven, and Thomas Villmann

There are at least three types of generalization in the broader context of cognitive science and artificial intelligence research:

1. Generalization refers to a process by which general concepts and rules are constructed from example data.
2. Generalization refers to the product of such a process, meaning the general concepts and rules themselves, in their diverse representations.
3. Generalization refers to the application of a product to new data.

The types of Generalization group dove deep into these three types and their sub-types, drawing on prior work from cognitive science, symbolic artificial intelligence, and machine learning. Key theories of generalization deal with abstraction, adaptation, domain extension, composition, analogy/transfer, and in vs. out of distribution. Important representations of generalization are:

- symbolic rules,
- prototypes/exemplars,
- probabilistic distributions, and
- functional mappings.

The coordinator of this working group was Benjamin Paassen.

Participants

- Wael Abd-Almageed
Clemson University, US
- Michael Biehl
University of Groningen, NL
- Marianna Marcella Bolognesi
University of Bologna, IT
- Xin Luna Dong
Meta Reality Labs – Bellevue, US
- Kenneth D. Forbus
Northwestern University –
Evanston, US
- Kiril Gashteovski
NEC Laboratories Europe –
Heidelberg, DE
- Barbara Hammer
Universität Bielefeld, DE
- Pascal Hitzler
Kansas State University –
Manhattan, US
- Filip Ilievski
VU Amsterdam, NL
- Giuseppe Marra
KU Leuven, BE
- Pasquale Minervini
University of Edinburgh, GB
- Martin Mundt
TU Darmstadt, DE
- Axel-Cyrille Ngonga Ngomo
Universität Paderborn, DE
- Alessandro Oltramari
Carnegie Bosch Institute –
Pittsburgh, US
- Benjamin Paaßen
Universität Bielefeld, DE
- Gabriella Pasi
University of Milan, IT
- Sascha Saralajew
NEC Laboratories Europe –
Heidelberg, DE
- Zeynep G. Saribatur
TU Wien, AT
- Ute Schmid
Universität Bamberg, DE
- Luciano Serafini
Bruno Kessler Foundation –
Trento, IT
- Dafna Shahaf
The Hebrew University of
Jerusalem, IL
- John Shawe-Taylor
University College London, GB
- Vered Shwartz
University of British Columbia –
Vancouver, CA
- Gabriella Skitalinska
Leibniz Universität
Hannover, DE
- Clemens Stachl
Universität St. Gallen, CH
- Gido van de Ven
KU Leuven, BE
- Frank van Harmelen
VU Amsterdam, NL
- Thomas Villmann
Hochschule Mittweida, DE
- Piek Vossen
VU Amsterdam, NL
- Michael R. Waldmann
Universität Göttingen, DE

