Artificial Intelligence Research Agenda for the Netherlands

van den Bosch, A.; van Dijck, J.; Helberger, N.; Heylen, D.; Hindriks, K.; Hoos, H.; Lagendijk, I.; de Rijke, M.; Niessen, W.; Verheij, B.; Vossen, P.; van Wynsberghe, A.

Publication date
2019

Document Version
Final published version

AIREA-NL

Artificial Intelligence Research Agenda

for the Netherlands

NWO

COLOFON

Publishing date : 1 November 2019
Design: WAT ontwerpers
Published by NWO
Contact: airea.nl@nwo.nl
NWO is affiliated with the NL AI coalition

NL AI Coalition

Artificial Intelligence Research Agenda

for the Netherlands

EXECUTIVE SUMMARY

Artificial intelligence (AI) has evolved from a technology for niche applications to a transformative force that affects the welfare and wellbeing in many, if not all, societies. The introduction of AI technology – often using massive amounts of digital data and computing resources – in governmental processes, healthcare, industrial processes, the service industry, and many other sectors, has major economic and sociocultural impacts. The use of AI changes our lives, work, and organisations, and therefore affects the very fabric of our society. AI strategies are currently being defined and executed with high urgency worldwide to create opportunities for benefitting from AI, understand the implications of AI and create the right conditions for AI. However, a national strategy can only succeed if we have a fundamental knowledge of and control over these AI technologies. Furthermore, governments can only steer AI towards societal good and acceptance if they have the means and technological knowledge to act effectively.

The Netherlands has recently launched its national AI strategy. This strategy builds on the vision that AI systems should aim at effective collaborations between humans and AI systems, rather than at replacing human skills and capabilities. The AIREA-NL AI research agenda is an integral part of that strategy. It describes the central challenges to be addressed by researchers in the Netherlands so that the Netherlands can fulfil its ambition of being a European leader in AI technology and contributing to the shared European vision of human-centred AI.

This AI research agenda is organised around grand challenges related to different facets of the life cycle of an AI algorithm. Each grand challenge recognises how technological and societal aspects of AI are inherently intertwined and mutually influence each other. This agenda is the convergent result of contributions, discussions and feedback from many researchers in a wide range of disciplines who through relevant research, contribute to responsible AI technology and applications. Taking a perspective on AI as a socio-technical system, we encourage collaborations between different strands of the natural sciences, technology, social sciences, and humanities.

The four specific facets of AI grand challenges addressed in this research agenda are: creating AI components, creating AI systems, AI systems and

humans, and AI systems and society. Three research challenges are articulated for each facet. The research agenda also identifies communal and often multidisciplinary cross-cutting questions: responsibility and accountability, explainability and transparency, human alignment and social awareness, generalisability and contextualisation, and data and energy efficiency. These cross-cutting considerations ensure close alignment with the European vision of human-centred AI.

The fundamental and applied research results from this AI research agenda will push back the boundaries of our AI knowledge. The broad benefits of these scientific advances can only be realised if there is intensive collaboration between fundamental and applied AI researchers, practitioners and application domain experts. The agenda comprises a selection of such application domains with an emphasis on the common good; a selection that is well-aligned with the Dutch, European and UN Sustainable Development Goal challenges, with sociocultural and industrial impact, and with scientific discoveries.

The Dutch AI research agenda builds on three decades of sustained and internationally recognised AI research and education and aims to position the Netherlands among the world leaders in AI to maximise the benefits of AI for Dutch welfare and wellbeing. Full commitment and investment in well-funded fundamental and applied research are essential. That is the only way the Netherlands can attract and retain the necessary talent, and be at the forefront of research, innovation and industrial and societal applications of artificial intelligence.

Editors

Expert Committee

Thomas Bäck, Leiden University
Sander Bohté, CWI
Peter Boncz, CWI
Tibor Bosse, Radboud University
Philip Brey, University of Twente
Mehdi Dastani, Utrecht University
Francien Dechesne, Leiden University
Guszti Eiben, VU Amsterdam
Janneke Gerards, Utrecht University
Bram van Ginniken, Radboud University
Maaike Harbers, Rotterdam University of Applied Sciences
Pim Haselager, Radboud University
Tom Heskes, Radboud University
Geert-Jan Houben, TU Delft
Marleen Huysman, VU Amsterdam
Franciska de Jong, Utrecht University
Uzay Kaymak, TU Eindhoven
Max Louwerse, University of Tilburg
Gertjan van Noord, University of Groningen
Nanda Piersma, Amsterdam University of Applied Sciences
Eric Postma, University of Tilburg
Hedderik van Rijn, University of Groningen
Johannes Schmidt Hieber, University of Twente
Cees Snoek, University of Amsterdam
Matthijs Spaan, TU Delft
Niels Taatgen, University of Groningen
Suzan Verberne, Leiden University
Wijnand IJsselsteijn, TU Eindhoven
Harry van Zanten, University of Amsterdam

Editorial Committee

Antal van den Bosch, KNAW Meertens
José van Dijck, Utrecht University
Natali Helberger, University of Amsterdam
Dirk Heylen, University of Twente
Koen Hindriks, VU Amsterdam
Holger Hoos, Leiden University
Inald Lagendijk, TU Delft
Wiro Niessen, Erasmus MC/TU Delft
Maarten de Rijke, University of Amsterdam
Bart Verheij, University of Groningen
Piek Vossen, VU Amsterdam
Aimee van Wynsberghe, TU Delft

NWO Support

Paul Blank, NWO Applied and Engineering Sciences
Frank Karelse, Taskforce for Applied Research SIA
Janneke van Kersen, NWO Social Sciences and Humanities
Marlies van de Meent, NWO Social Sciences and Humanities
Astrid Zuurbier, NWO Science

PREFACE

The AIREA-NL agenda is the result of contributions by many AI researchers in the Netherlands across the natural sciences, applied and engineering sciences, social sciences and humanities. The AI Manifesto of IPN's Special Interest Group on AI was a useful starting point in the process of putting together the AIREA-NL agenda. SIG-AI represents all computing science academic institutes and researchers in the Netherlands that perform AI research. For many of the computer science challenges in Section 4, the AI Manifesto provides a far more extensive description. The expert committee provided input for the urgency, scope and initial structure and grand challenges of the AIREA-NL agenda, as well as the challenges in specific AI areas. The editorial committee first consolidated these inputs into the grand challenges, research questions and cross-cutting considerations. The editorial committee then jointly wrote the entire draft version of the research agenda. In an open public consultation process, a broad group of stakeholders across science, society and industry gave feedback on the draft version of the agenda. The editorial committee has made useful use of this feedback in composing the final version of the AIREA-NL research agenda. The entire process was greatly supported by NWO senior staff members.

CONTENT

1. URGENCY AND SCOPE

For many years the Netherlands has been among the world leaders in artificial intelligence (AI). It has contributed to many areas of AI, such as machine learning, automated reasoning, and multi-agent systems. The enormous recent success of data-driven AI in many societal and economic sectors is attracting large numbers of researchers worldwide to very well-funded AI research and development programs and institutions. A battle on talent is raging, at all levels, in science and industry. Worryingly, academic AI research in the Netherlands is rapidly losing ground and attractiveness, due to increasingly non-competitive academic conditions, and partly due to ambitious and well-funded AI initiatives in neighbouring countries. This AI Research Agenda for the Netherlands (AIREA-NL) outlines fundamental challenges in which the Netherlands must urgently invest if it has the ambition to remain among the AI science and technology leaders, and thus be able to exploit future economic and societal benefits of AI.

Today's AI is the science and engineering of making machines intelligent and collaborative. The research field was founded in the 1950s, and since then many foundational advances and engineering solutions have enabled machines to assist in tasks that require intelligence, such as reasoning, learning, finding information, understanding text, speech and images, listening and speaking in dialogue systems, and optimising complex systems. More recently, the availability and (re-)use of massive amounts of data, the growth of computational power, and the optimisation of algorithms have resulted in an enormous leap in performance and versatility in AI algorithms. These AI technologies are now beginning to penetrate society at an unprecedented scale, and the deployment and use of AI algorithms has become ubiquitous. AI is a factor of major importance in every country's future prosperity and well-being. New challenges are consequently emerging, such as social collaboration, societal integration, explainability and responsible use. AI is here to stay.

To benefit from advances in artificial intelligence, AI strategies are currently being defined and executed worldwide.
The urgency to do so stems from five reasons unique to AI compared to other technological advances[1]. First, AI is a

1. https://www.vno-ncw.nl/sites/default/files/position_paper_algoritmen_die_werken_voor_iedereen.pdf

winner-take-all technology, because AI algorithms successfully trained on data attract more usage, more data, hence result in better performance and economic advantages that can be leveraged. Second, AI is a chess piece in the geopolitical race for power, talent and economic supremacy. Competing continents have different policies for AI development and deployment; Europe strongly emphasises the need for human-centric AI and AI that respects fundamental rights. Third, the rapid expansion of AI R&D has caused a worldwide pursuit for the best talent. AI talent is attracted by the presence of talent and by a healthy AI ecosystem in terms of research, innovation, infrastructure, and societal readiness. Fourth, acceptance of AI technology as a societal force is not always uncontested, hence mitigating risks becomes a necessity. The worldwide discussions about the impact of algorithmic systems on human rights and the prohibition of autonomous lethal weapons are two relevant examples. And finally, AI will create massive shifts in the job market, and upskilling/reskilling of many categories of employees must begin now.

Recognising these strategic urgencies will be to no avail if a country does not have a technology leadership position. A country can only succeed in critical societal and economic sectors if it has fundamental knowledge of and control over AI technology. It can only steer AI towards societal good and acceptance if it is an actor in the technology development itself. Undesirable dependencies on countries and companies not sharing our values must be avoided. A seat at the international AI table demands independent, world-class research at the national level. And it is also key to driving the innovation in applied research, government, industry, SME and start/scale-ups. For this reason, AI strategies of technologically advanced countries all have the explicit ambition to be a technology leader, and they all underpin this ambition with a significant academic research programme.

The Netherlands recently released its nationwide AI strategy[2,3]. Ambitious and well-funded AI research must be a key ingredient of that strategy. This AI research agenda describes the fundamental challenges that need to be addressed for the Netherlands to fulfil its ambition to be among the European AI technology leaders and contribute to the European vision of human-centred AI. Addressing these challenges is essential if the Netherlands wishes to see AI contribute to societal challenges and the UN Sustainable Development Goals, and also paramount to seamlessly integrate AI in key Dutch economic sectors and vital infrastructures (for instance, agri-food, energy, life science & health, logistics, safety and defence) and in the Dutch sociocultural fabric.

Nurturing AI research in a healthy ecosystem is important. First and foremost, education of highly skilled AI researchers

2. Government AI strategy SAPAI: https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/kamerbrief-ai.
3. The Netherlands AI Coalition: https://nlaic.com/.

and practitioners across all disciplines and sectors is essential because continuous integration, adaptation and improvements within local context demands specialists within the Netherlands. Second, academic talent at all seniority levels must be retained, for instance by offering competitive starting packages for junior researchers, and interesting (collaborative) projects for everyone to work on. And third, results must be disseminated according to the Open Science principles, which is not trivial when proprietary datasets are used for learning AI systems, and the winner-take-all economic model dominates. These aspects should all be part of the overall Dutch AI strategy. The scope and urgency of this AIREA-NL agenda is to fill the important gap in the national investment landscape by focusing on the fundamental research needed for future leadership in AI technology that is embedded in society in a responsible manner. The agenda aims at a balance between the different perspectives on what is needed for making AI an effective and impactful technology for the common good; much of what will undoubtedly follow will require multidisciplinary collaboration.

2. MOTIVATION AND STRUCTURE

The field of AI has gone through different cycles since its inception. The work of Turing and contemporaries laid a solid theoretical foundation for AI as a computing discipline. Different waves of progress brought the foundations and successes of neural networks, rule-based AI systems, and statistical machine learning techniques. Throughout much of its history, AI has fruitfully interfaced with neighbouring disciplines, including mathematics, cognitive science, and human-computer interaction. The Netherlands has a long-standing tradition in contributing to these developments in AI research and developing AI-related education programmes at the BSc, MSc and PhD levels[4]. The Netherlands is home to some of the world-leading groups in social sciences and humanities (SSH), doing research on the legal, ethical and sociocultural dimensions of technology. Over the years, extensive data research infrastructures have been set up in the Netherlands, which are facilitating scientific research in different domains.

But AI today is very different from AI a decade ago. For a long time, advances in AI that resulted from fundamental algorithmic research had limited performance and

4. https://www.nwo.nl/documents/enw/rapport-ai-voor-nederland-vergroten-versnellen-en-verbinden

narrow impact. Only ten years ago, when the computing and data revolution gained speed, did machine learning-based AI take off and AI start to show qualified performance and versatile applicability. Software libraries such as TensorFlow and PyTorch now make machine learning-based AI available to any scientific discipline and to any organisation for integration in their workflow. The Dutch AI community at large, from multi-agent to machine learning researchers, has greatly contributed to this development, as evidenced by several reputed Dutch academics and the founding of a number of highly successful AI start-ups from universities. These researchers are ideally positioned to help the Netherlands in maintaining a leadership position in selected technology areas.
Because AI systems will have a significant impact on society, the call for human-centred AI and a societal perspective on AI development is growing louder from industry, academics, governments and multi-stakeholder organisations. Technology does not operate in isolation but in the context of existing societal, cultural, and institutional practices and norms. The successful and widespread deployment of AI technology hinges on sociocultural compatibility and acceptance. Conversely, AI will impact societal, ethical, legal and economic factors, and in this way change society. Thus, requirements need to be identified that enable us to make such changes for the better and avoid unintended consequences. The social compatibility of AI requires constant study and evaluation as AI unfolds across society.

Research and development of today's AI systems are increasingly aiming at effective collaborations between humans and AI systems, rather than merely at replacing human skills and capabilities, so that collaborative solution strategies can be

developed that outperform purely human and purely AI-based strategies. This type of hybrid intelligence emphasises the multidisciplinary nature of the field. It introduces cross-cutting requirements on AI algorithms that were mostly irrelevant in the early days. Examples of such cross-cutting considerations are that AI algorithms are explainable and responsible, and that AI techniques are fair, data efficient and can generalise over application domains. Many of these challenges have only appeared in recent years and are becoming the focus of the next wave of AI technology, leading to AI algorithms that augment human intelligence.

An AI research agenda that describes challenges vital for being among the European leaders in AI technology will need to reach far beyond today's state-of-the-art and practice. It will require solving urgent fundamental and technological AI problems in combination with addressing societal implications and constraints. Fundamental research in the Netherlands should embrace the strategic direction Europe has chosen to aim for human-centric AI. We aim for European AI in the Netherlands.

FACETS OF AI ALGORITHMS

We have chosen to organise what we consider to be the AI grand research challenges according to four specific facets of an AI algorithm's life cycle (see Figure 1). The grand challenges recognise how the technological and societal aspects are inherently intertwined; the social aspects impact the technical ones and vice versa. The result of this is understanding AI as a socio-technical system.

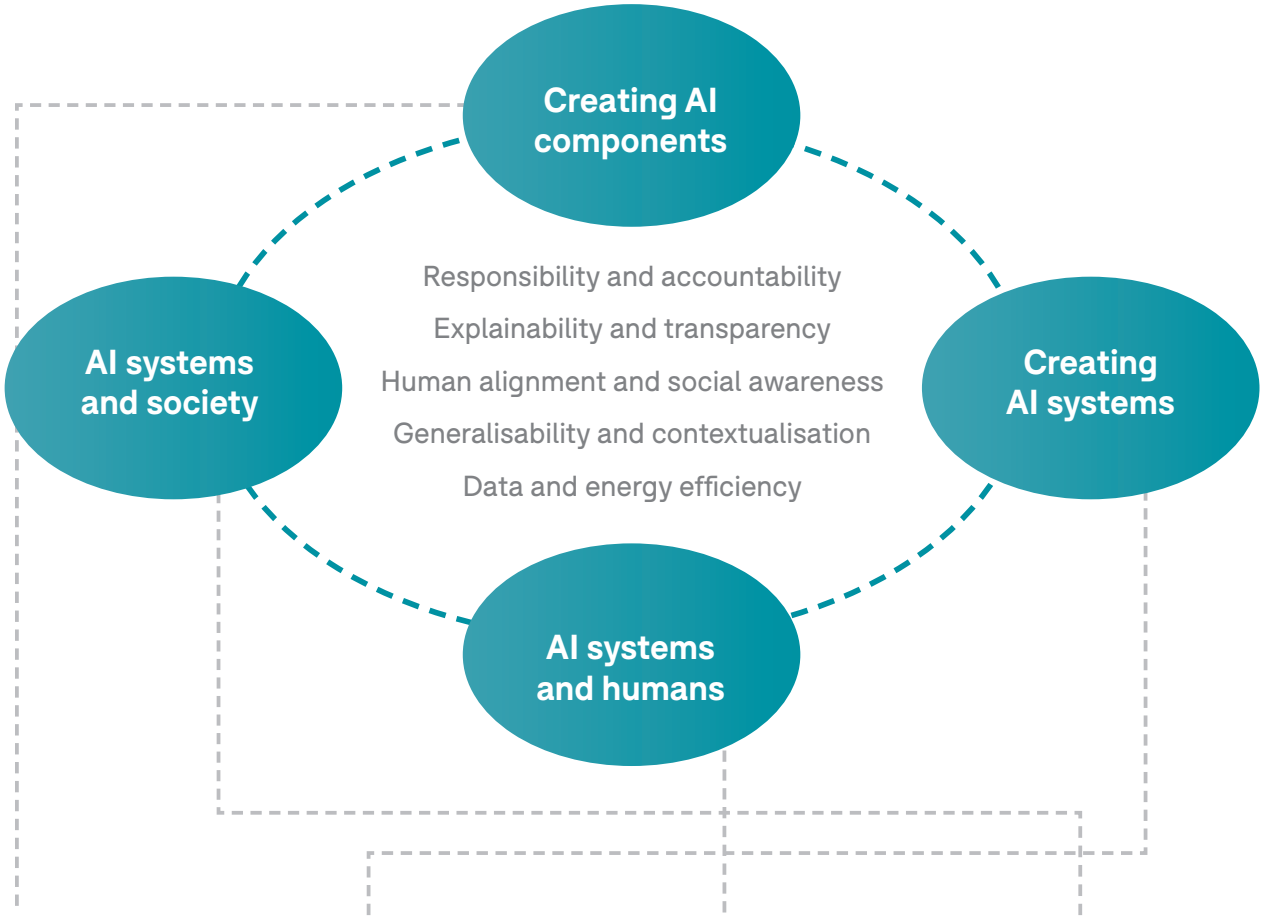
The first facet is "creating AI components". This involves research into new AI algorithms, embedding the relevant performance, human, societal and cross-cutting factors. The second facet is "creating AI systems". Here, the focus is on getting systems built and used, systematic methodologies, the interaction of different AI components, the predictability of the overall behaviour of the system within a particular usage context, and the access to data and knowledge. The third facet addresses how "AI systems and humans" can learn from each other and optimally collaborate, including aspects particular to the Netherlands such as language. Finally, the fourth facet concerns "AI systems and society" and addresses how the transformative force of AI and society interact and shape each other. For each of the four facets we will formulate three central research questions to be addressed.

CROSS-CUTTING CONSIDERATIONS

Each of the four facets of AI algorithms has its own fundamental challenges from the perspective of technology innovation and responsible societal embedding. At the same time, similar urgent and often multi-disciplinary questions anchored to the unique properties of AI are present across these facets. The socio-technical nature of AI means that these questions are cross-cutting and cannot be put into a single facet

of AI. In fact, the European vision towards human-centred AI aligns well with the five cross-cutting considerations that we highlight in AIREA-NL. To emphasise that we subscribe to the European agenda, we have highlighted the multidisciplinary cross-cutting considerations in a set of framed boxes, focusing on "responsibility and accountability", "explainability and transparency", "human alignment and social awareness", "generalisability and contextualisation", and "data and energy efficiency".

Figure 1: Overview of the facets and cross-cutting considerations of the grand challenges in AIREA-NL.



RQ-1.1. How can AI algorithms learn continuously and robustly?

RQ-1.2. How to combine data-driven learning and model-based reasoning?

RQ-1.3. How can we design trustworthy AI components with appropriate quality assurances?

RQ-2.1. How to lower the barrier to implementable AI using a systematic design approach?

RQ-2.2. How to create a knowledge and data infrastructure that is maximally conducive to the construction of trustworthy and effective AI systems?

RQ-2.3. How to predict the overall behaviour and quality of AI systems?

RQ-3.1. How can humans and AI systems productively interact and understand each other's behaviour in context?

RQ-3.2. How can we realise AI systems that deserve human trust?

RQ-3.3. How can we design tasks for hybrid human and AI teams?

RQ-4.1. How do we ensure that everyone benefits from AI?

RQ-4.2. How do we prevent the design and use of AI systems from perpetuating unwanted histories and unacceptable behaviour?

RQ-4.3. How do we design value-sensitive, norm-aware AI systems?

3. GRAND AI RESEARCH CHALLENGES

3.1 CREATING AI COMPONENTS

Progress in AI in the past decade has, to a large extent, come from new insights in data-driven approaches. However, the significant performance improvements thanks to these approaches are starting to see their limits. We identify three limiting factors which require the research into AI components. Each of these components has AI technology and SSH perspectives.

Research Question 1.1. How can AI algorithms learn continuously and robustly?

Most of today's AI systems are trained off-line with data at scale before they are operationalised. But large volumes of data are often not available, because examples of the phenomena to learn are scarce or because it is not feasible to collect large-scale (labelled) data for legal, ethical, or technical reasons. How can we make machine learning more robust in using transfer learning, weakly labelled or unlabelled data, augmented data, and combinations of heterogeneous data such as speech, text, and images? Which data selection or augmentation strategy is optimal, not only to achieve efficient learning, but also to curtail the potentially sensitive exploration of the entity (human, machine, organisation) that is providing the learning examples? How can AI algorithms learn continuously while in use, and thus become lifelong learners, evolving and adapting their behaviour over time? How does long-term adaptivity of AI influence the design of AI component? How to deal with sudden shifts in statistics (for instance, in the context of natural language processing, in vocabulary used), or situations where the quality of the data or of its labels deteriorates, users of the system inadvertently provide incorrectly labelled examples, or adversaries deliberately attempt to throw off the AI algorithm? How can we validate, and provide safety guarantees for the performance and reliability of a continuously learning algorithm?

Research Question 1.2. How to combine data-driven learning and model-based reasoning?

Knowledge may be captured in patterns or in models. For human-level intelligence, data-driven and model-based approaches should be combined. How can machine learning use expert and world knowledge? How do we translate back and forth between the data-driven and model-based paradigms? How should ethical, legal, social requirements on AI algorithms behaviour

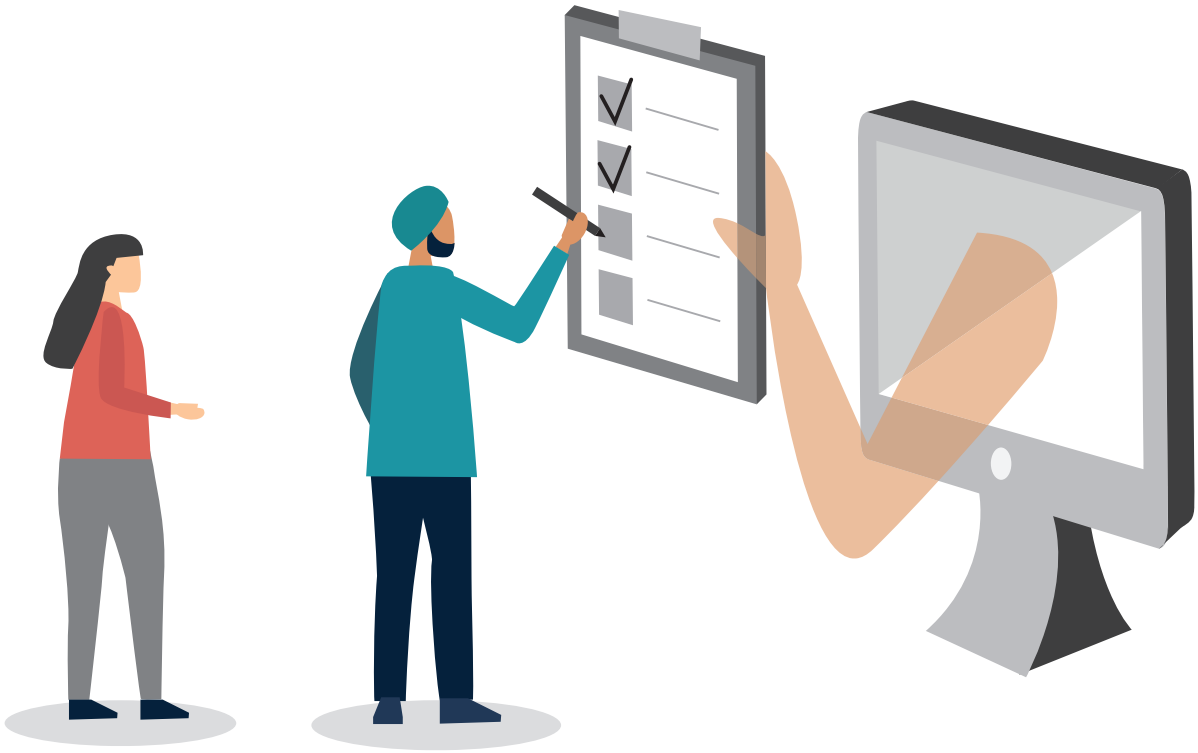
be captured in a data-driven paradigm? How can large-scale model-based approaches to complex decision-making be anchored in data collected through real-world experience? If AI algorithms are to make recommendations and decisions under constraints, then such constraints must be extracted from real-world use cases in a form amenable to combination with data-driven machine learning. And, conversely, how can global explanations (of the workings of a data-driven algorithm) and local explanations (of individual predictions by a data-driven algorithm) be generated?

Research Question 1.3. How can we design trustworthy AI components with appropriate quality assurances?

Today's artificial intelligence components are successful in their respective areas of AI (such as learning, reasoning, planning, search, information retrieval, vision, robotics), often using different measures of quality. For automated reasoning methods, quality assurances are focussed on logical correctness and time complexity. For learning and information retrieval, empirical quality measures are analysed statistically to establish component quality. Quality assurance lies at the heart of establishing AI methods and components as trustworthy, and many problems remain in this area, including scalability (how to ensure quality is maintained as the amount of data increases), and robustness (how to ensure that performance does not drop sharply and unexpectedly as the nature of given input data changes). Can different styles of quality assurances be combined? Can scalability be achieved without compromising robustness and vice versa? What mathematical foundations and methodologies can support quality assurances across a broad range of AI techniques? What software quality parameters should be considered, and how can AI components have safeguards and fall-back plans in case of problems or attacks? How can ethical, legal and social considerations be inserted in the quality assurances that AI algorithms should meet?

3.2 CREATING AI SYSTEMS

Whereas AI algorithms have made great progress in controlled and experimental settings, developing and applying an AI system “for use in the wild” is still far from trivial. After all, AI will eventually have to operate as a reliable autonomous (decision-making and acting) system. Design choices must be made, components of the AI system must be selected and configured, high-quality knowledge and data must be readily available, parameters must be tuned, the resulting system must be tested, and predictions of the overall behaviour are often required. As even the best of today’s AI systems will not give 100% perfect results, the system must be aware of its limitations and act accordingly, and comprise measures to deal with, for instance, adversarial conditions (such as deliberate attacks) and situations beyond those considered when building and testing the system. As AI systems are not only built up from AI components but also from data, we need to assure AI algorithms and data optimally fit together.



Research Question 2.1. How to lower the barrier to implementable AI using a systematic design approach?

The implementation of an actual AI system for a particular application requires the combination of different domain-independent AI components, such as model-based reasoning, machine learning, optimisation, computer vision and natural language processing. It also requires domain knowledge to be incorporated such as social, legal and economic structures of a particular sector. Today, choosing the right combination is still a bit of an art, and usually a divide-and-conquer (if not to say ad-hoc and trial-and-error) design principle is applied. How can we arrive at a systematic engineering approach for AI systems in different settings, such as centralised, distributed, embedded or resource constrained? Such a methodology would address questions such as: which components perform best under which circumstances? Can we design for a certain performance within bounded uncertainty? What are the legal, ethical, socioeconomic characteristics of the domain in which the AI system will need to operate? How can an AI architecture be designed to be computationally efficient and robust? Importantly, any such approach also needs to incorporate possibilities to elucidate and co-design for domain-dependent responsibility levels and explainability, so as to ensure that it produces trustworthy AI systems operating in a lawful, fair and robust fashion.

Research Question 2.2. How to create a knowledge and data infrastructure that is maximally conducive to the construction of trustworthy and effective AI systems?

Without access to real-world expert and institutional knowledge, and without access to real-world data, AI systems cannot successfully be designed and deployed. In an

optimally transparent world, this requires data collection and storage platforms compliant with the FAIR principles. Furthermore, it requires models that capture the complexity of the real world. For instance, much of the data in the Netherlands is distributed over different platforms, each subject to specific IP, informed consent, GDPR and other data protection regulations. Along the same lines, access to implemented AI algorithms and tools will accelerate the construction of AI systems. How can a federated AI approach be put in place, assisted by the appropriate contractual and technological measures to give AI systems responsible and compliant access to data, algorithms and tools? How do we describe dataset properties that enable effective processing by AI components in AI systems? How do we create data processing components that allow to guarantee desired dataset properties? Moreover, AI is sensitive to context, language and culture. Therefore, how can data, models and tools for Dutch culture and language be made available to designers of AI systems so as to support responsiveness to the requirements of the Dutch society?

Research Question 2.3. How to predict the overall behaviour and quality of AI systems?

With the increasing intelligence that AI brings to applications, how can we understand and predict the behaviour of AI systems in specific contexts? Or the behaviour of complex teams of humans and AI systems? How can we mathematically predict when failure is likely and when such failure is due to technological or human factors? How to understand what the quality and reliability of the AI system's actions are if it is fed with data of poor quality, possibly inconsistent with other information in the system? How do fairness, accountability, and transparency of AI algorithms or components translate into fairness, accountability and transparency of systems and organisational contexts in which these systems are used? How can we certify that the software for interacting and continuously learning systems can be verified, validated, and maintained over a long period of operation?

3.3 AI SYSTEMS AND HUMANS

Even though AI algorithms are still far away from human intelligence, we already see specific tasks where AI outperforms people. Embracing the vision of human-centred AI in which AI augments rather than replaces human intelligence, we aim for humans and AI to cooperate, leveraging complementary strengths, and where each agent features the tasks they are optimally suited for. For such hybrid AI to be successful, humans and machines must understand and trust each other, and new ways of collaborative problem solving are needed.

Research Question 3.1. How can humans and AI systems productively interact and understand each other's behaviour in context?

For an AI system to collaborate and coordinate with humans, it needs to be able to observe and understand, at least to some degree, human behaviour. And vice versa, how can AI explain itself so that humans

understand AI. How can machines capture and reason about the human state, such as dialogue and emotional state, information need, and interaction behaviour? How can AI systems and their designers understand the workplace or sociocultural context and their implicit cues? How do humans interpret a machine's predictions and how are they able to detect its errors? How can humans and AI systems be enabled to explain why a certain action was taken? What are specific human versus technological strengths, and how do we design the right conditions and incentives so that they optimally cooperate and complement each other? And what are the long-term effects of such interactions on AI system development, human behaviour, skills and society? How are human cognitive abilities impacted? And how, in turn, do we design short-term and long-term rewards to make sure AI systems behave as desired?

Research Question 3.2. How can we realise AI systems that deserve human trust?

If a human collaborator or user understands the behaviour of an AI system, this does not automatically mean that the AI system is trusted or should be trusted. And similarly, AI systems need to be critical towards themselves and the people with whom they collaborate and from which they get input. How are trust relations with AI systems formed, valued and evaluated? How do we prevent humans from overendowing AI systems with "real" understanding on the one hand, possibly leading to overly optimistic trust, and from algorithm aversion on the other hand, possibly leading to suboptimal decision-making? Different interaction modalities come with different challenges, including human decision-making where AI systems recommend, machine decision-making where a human is subjected to an AI system, and hybrid forms where decision-support and decision-making are mixed. How do we design actionable interventions, checks and balances, and contestability for people and systems in all these decision-making scenarios? How do we achieve meaningful human control of AI systems? And how can cognitive constraints and ethical, legal and social expectations be translated into requirements for people interacting with AI systems, and for operationalisation of those systems?

Research Question 3.3. How can we design tasks for hybrid human and AI teams?

There is growing evidence that hybrid human and AI teams can develop solution strategies that neither human nor AI teams by themselves would have produced. How do we recognise such tasks? How do we organise human-AI system collaborations so that better decisions emerge? The future of hybrid AI-human work will require a new set of skills; not just technological skills such as programming, but higher cognitive skills such as creativity, critical thinking and complex information processing should be paired with social and emotional skills. How can we optimise learning to promote such diverse skills and orientations? What new skills will need to be acquired? How do we incentivise hybrid teams, make them diverse and ensure a smart division of tasks between humans/AI, and optimal working conditions for human workers? Should AI systems that

form part of such hybrid teams be optimised for tasks at which they outperform humans (e.g., number crunching or large-scale analysis of data) or for human cognitive tasks to accelerate the development of innovative strategies? How will hybrid AI affect the future of work?

3.4 AI SYSTEMS AND SOCIETY

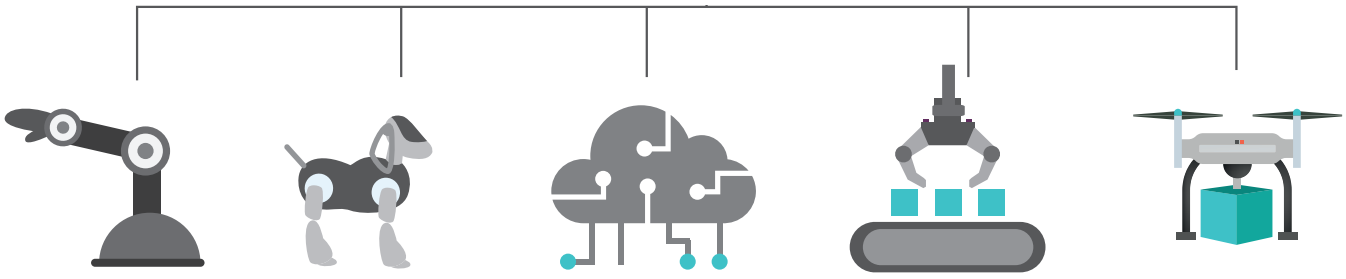
In the long run, technology leadership in AI is a means, not a goal in itself; it must serve human progress and welfare. In many circumstances, the ability to make predictions in the face of (data about) many uncertain factors is very useful. This ability can give great power, and with power comes responsibility. Hence, for humans and society to collectively benefit from AI technology requires integrating AI into society in a way that respects fundamental rights, public values and social dynamics.

Research Question 4.1. How do we ensure that everyone benefits from AI?

Today, a relatively small subpopulation of highly-educated and skilled, overwhelmingly male and well-paid workers determine how AI systems are designed, and what purposes they are optimised for. By the same token, the economic benefits of AI may not be distributed equally if the gap between low and high-skilled workers increases. How can we capture attitudes, needs, and assumptions to characterise AI systems and their use beyond the group of technology developers involved in today's design, implementation and deployment trajectories? How can we recognise and address unintended consequences of the lack of diversity in AI research and development? How can we even predict who will be disadvantaged by (the lack of) specific AI properties or design and act against discrimination and the new digital inequality? How can we broaden the base of those that benefit from AI systems and ensure the economic impact of AI technologies gets distributed evenly between societal groups? For AI to benefit different sectors - each with its own needs, economics, norms and practices - and groups of stakeholders, it is necessary to understand how to operationalise these characteristics in the design of AI components, systems and implementation. How can we adjust systems of governance and regulatory principles to ensure a fair socioeconomic implementation of AI-based applications?

Research Question 4.2. How do we prevent the design and use of AI systems from perpetuating unwanted histories and unacceptable behaviour?

AI systems, especially those that fall under the category of supervised learning, reflect what is already given, and not what could or should be. Biases and other harmful tendencies in society will be present in datasets. Whenever patterns are extracted using machine learning and used for forecasts and predictions about unknown attributes, there is a risk to perpetuate these flaws, and to thus contribute towards faulty system behaviour, or confirming the status quo of undesirable social conditions, cultural oppositions, and power structures. How can AI systems generate solid



predictions while producing creative, culturally and socially acceptable outcomes? How do we codify and implement such desired behaviour? What are actionable interventions to bring about changes, and which rights should users have? The effects of AI do not only depend on the technology itself but also on the way is integrated in society and who controls it. We need a clear understanding of the societal dynamics of AI, the broader economic and political power structures behind AI and how a transition to AI affects humans and society. To that end we also need a better understanding of the broader macroeconomic consequences of the growing concentration of societal and economic power in a few players, and the implications for independent innovation, functioning competition and regulation.

Research Question 4.3. How do we design value-sensitive, norm-aware AI systems?

AI systems and the underlying architectures are not developed by simply using a particular program and applying it in a given use context. AI systems and their interactions need to be adapted to the specific dataset and context in which they are used. In the design of a data-driven AI system, how can we judge if a particular set of training data does indeed match the requirements for the system being developed? As the complexity of interacting AI systems continues to grow, to what degree should their design be automated? And how can we ingrain legal, ethical and social qualities, such as fairness, accountability, and transparency, into an automated design process and its outcome? What are the values to optimise for, and how do these values change under the impact of AI? How can we ensure human agency, meaningful human control and oversight in the development, deployment and use of AI? Can design for values approaches express critical trade-offs in AI which are the result of shifting powers and conflicting (economic) incentives and power structures? What safeguards are needed in situations in which values do not lend themselves to easy implementation? And how to design for, and implement, resilient AI systems, incorporating potential vulnerabilities to attacks that may throw off or disable a mission-critical AI system?

Cross-cutting consideration I: Responsibility and Accountability

Even more so than simpler computational approaches, AI is prone to "garbage in, garbage out". Any data, information or knowledge on which AI technology is based or from which it learns is somehow biased, because of the moment it was gathered, by which process, for which purpose. This bias may be overt, but is more often hidden, underestimated, or simply unknown. Yet, if your skin colour bars you from using face recognition software, or when your dialect is not recognised by the generic speech recogniser made for the standard language, you realise as a user that decisions were made, consciously or not, that resulted in discriminatory performance. In all facets of AI and AI use, it is vital to acknowledge the responsibility to remove bias and to make sure corrective actions can be taken if unintended consequences occur in order to respect inclusiveness, non-discrimination, fairness, privacy, autonomy, dignity, accountability, and due diligence. To that end, we need to be able to steer AI and its implementation normatively and lawfully, such that outcomes meet our responsibility criteria, create effective interventions and accountability. This requires the balancing of objective descriptive aims, and subjective normative aims. It also requires effective division of responsibilities between the different actors as, often, processes of algorithmic decision-making do not always have a single "owner" or "applying actor". New frameworks are needed that can guide us in identifying the responsibility issues and help us to ensure effective (human) agency, accountability and oversight.

Cross-cutting consideration II: Explainability and Transparency

As AI system's actions and decisions will significantly affect their users, it is important to be able to understand how and why an AI system produced the effect that it did. It is a well-known hurdle that many AI algorithms behave largely as black boxes. For example, predictors obtained from state-of-the-art deep learning techniques often perform well in terms of the input-output function they represent but are hard to make sense of. It becomes even harder to understand the action of a system that is composed of multiple interacting AI components and possibly also human agents. The first aim of explainability is therefore to make the inner workings of AI systems more accessible and transparent. Secondly, it requires techniques to make causal and rational models explicit so as to create satisfactory explanations that are intelligible for human users interacting with the system. Conflicting with explainability may be domain-specific requirements that demand keeping the exact workings of an algorithm or system secret, for instance because of commercial stakes or national security risks. This necessitates a balance to be struck between explainability and the public and commercial interests involved. There may also be complex trade-offs between explainability and performance. Finally, explainability must be actionable, resulting in concrete interventions, where needed, to safeguard fundamental values and rights of individuals.

Cross-cutting consideration III: Human Alignment and Social Awareness

For an explainable, responsible and social AI, aiming for human alignment and social awareness is necessary and suggests valuable design hypotheses. Since the start of the field, results and methods from cognitive science and the social sciences have been a part of AI developments across the field. Knowledge has been represented as rules, scenarios and frames, allowing for both human and machine readability. Reasoning in AI has been styled according to the logical, statistical and critical methods known from rational and empirical methodology used in all sciences. Learning techniques have been inspired by the neural structure of the brain. Natural language processing remains an important component for a human aligned, socially aware AI. Human-agent interaction is an important topic in AI in these days of interconnectedness and internet of things. This field builds on and requires further development of socially aware AI techniques. Key challenges are whether human alignment and social awareness is best modelled based on human-human interaction, or whether artificial methods can achieve better outcomes. And how can AI technology and the humanities and social sciences co-create effective and socially aware AI? How can we inform people and explain why automated decisions are made by an AI system? How can we balance tasks between humans and AI component or systems, such that an optimal balance is obtained, taking the strengths and limitations of human perception and cognition into account?

Cross-cutting consideration IV: Generalisability and Contextualisation

Many state-of-the-art AI systems involve learning from a set of (labelled) training data. Provided that enough training data are available, these techniques are good at interpolation. They work well in conditions that are like those present in the training data. However, in practice, when AI is deployed in a real-life scenario, collaborating with real people in real organisations, it is likely that new conditions occur that are different from those present during the AI design. Generalisation to such new conditions and, indeed, the ambition to build all-purpose AI that is sensitive to the context and requirements of the actual systems into which AI is implemented, is a challenging endeavour. For instance, the application and optimisation of algorithmic personalisation competes with the idea of generalisability of algorithms. It becomes important to be able to assess whether a technology has been sufficiently trained and evaluated for the environment in which it is expected to operate. Adaptive training, transfer learning and auto-ML are potential approaches for AI systems to quickly adapt to new situations. At all levels of AI fall-back scenarios are important, which demands redundant system design and self-reporting on the confidence of AI recommendation and actions.

Cross-cutting consideration V: Data and Energy Efficiency

Modern AI systems based on machine learning provide highly scalable solutions for problems in computer vision, information retrieval, and natural language technology, all of which attain state-of-the-art performance when trained with large amounts of data. In these domains, the challenge we now face is how to learn, reason, perceive and communicate efficiently with the same performance in less time, with less data, and consuming less power. Other problem domains, such as autonomous driving, gesture recognition, personalised healthcare and robot learning are often characterised as small-data problems. The ability to learn, reason, perceive and communicate in a sample-efficient manner is a necessity in these data-limited domains. The need to design software and hardware energy-efficient techniques and architectures for machine learning, reasoning and perception is felt across the whole spectrum of computing systems – ranging from low-end mobile devices running at the edge to large-scale data centres and servers. Collectively, these problems highlight the increasing need for data- and energy-efficient AI: the ability to learn, reason and perceive in complex domains without requiring large quantities of data or energy. Like many other developments in society, AI must also aim for minimising any of its environmental impacts.

4. CHALLENGES IN SPECIFIC AREAS OF AI

Scientific and technological progress in many AI areas is needed to solve the grand research challenges and cross-cutting considerations. In this section, we concisely highlight important AI areas in which challenges exist that the Netherlands can contribute to, based on the strong reputation of academic research groups. We refer to IPN's SIG-AI Manifesto[5] for a more detailed description of the areas in Sections 4.1-4.7.

4.1 MACHINE LEARNING

Recent developments in machine learning, and in particular deep learning and reinforcement learning, have fuelled enormous progress in AI, leading to improved image and audio analysis tools, better machine translation, as well as better reasoning, planning and optimisation algorithms. However, current machine learning models do not understand the world yet at a level that humans do, and this lack of context makes it very hard for them to generalise to new situations. They also usually require large amounts of carefully curated training data. Humans achieve good performance at tasks within domains in which they can rely on background knowledge and understanding, also when there is little data. Here background knowledge may come in many forms, such as cultural and human traits, or physics of the world and objects around us. How can we integrate such knowledge into machine learning algorithms? As many situations in which AI will be applied do not comply with the prevalent paradigm of supervised machine learning, data acquisition for learning then depends on the system's own behaviour. How do we develop more efficient reinforcement learning algorithms, balancing exploration and exploitation of data and avoiding the danger of self-fulfilling prophecies? How do we obtain learning algorithms that are robust to deviations from key characteristics of the data that was used to train them? How can algorithms be trained on data from one specific use case be efficiently adapted to different use contexts?

4.2 KNOWLEDGE REPRESENTATION AND REASONING

Symbolic AI and logical methods form the basis for hard- and software verification,

5. http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf. See also Appendix A.

which are amongst the economically most impactful applications of AI to date. They are also becoming increasingly important in the context of the need for explainable and responsible AI, as both require reasoning about AI systems. To that end, notions such as agent concepts (including affective and intentional stances, e.g., belief, intention, ...), argumentation theory, knowledge representation and reasoning and ontologies must be made amenable to computerised treatment and processing. How can symbolic reasoning concepts, which are often qualitative knowledge-based methods, be integrated with today's successful data-driven methods, which are often purely numerical and statistical? What are robust representation and reasoning techniques for knowledge and data that is large, dynamic, heterogeneous and distributed? How can we make effective use of knowledge representation and reasoning techniques when addressing other AI challenges, such as vision, natural language understanding, question answering, and robotics? And how can we scale up existing automated reasoning techniques, such as SAT and SMT solvers, so they can verify important safety and correctness properties of complex software systems, such as AI systems?

4.3 PLANNING AND SEARCH

AI planning and search aims to provide algorithmic solutions for both single and multi-agent planning and scheduling problems, as well as for search and optimisation problems. This involves algorithms to efficiently search in general solution spaces, and methods specifically for planning and scheduling, sequential decision-making for one or more parties (multi-agent systems) under uncertainty, game-theorical approaches, adaptive decision strategies, mechanism design, social choice theory, and combinations of search and machine learning algorithms (as in, e.g., AlphaGo). It also aims to design and understand fundamental properties of methods to support intelligent decision-making. A key issue is reliable uncertainty quantification: a good decision-maker, whether human, machine or hybrid solution, needs to have an idea of the potential suboptimality of its decisions, for example, statistically valid bounds of the probability of bad outcomes. Decisions need to be socially aware, considering stakeholders and their preferences, and the (reasons for these) decisions need to be explainable to human experts and policy makers. Decision-making must be adaptive, as circumstances in which decisions must be made may change. Adaptivity is a requirement for an automated decision-making system to be self-correcting. Although there has been steady progress on algorithms to support automated decision-making, making these systems more effective and responsible is an important challenge.

4.4 COMPUTER VISION

The world is adapting swiftly to visual computing, communication and intelligence via the internet, mobile phones and platforms & devices equipped with cameras.

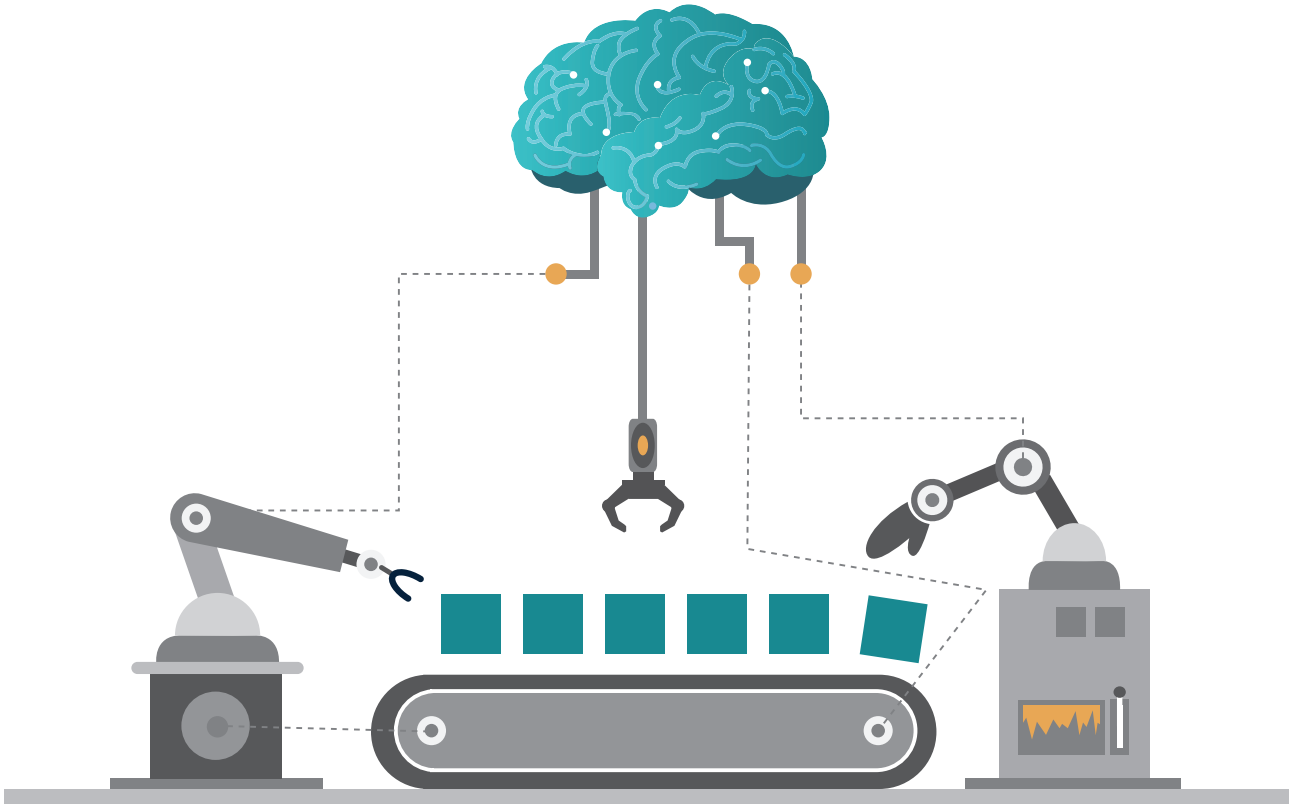
The super-human image classification performance achieved by deep learning in the ImageNet competition is the leading example to stress the breakthrough in this area. Nevertheless, automatically understanding the full complexity of visual content requires progress in colour processing, semantic understanding, 3D reconstruction, interactive picture analysis, image and video retrieval, human-behaviour analysis, and event recognition. What algorithms do we need for or visual interpretation based on precise appearance and geometry understanding? How can we design vision algorithms that require less expert supervision and generalise to novel visual domain? And how can we combine vision with techniques from machine learning, reasoning, natural language, and robotics?

4.5 NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) uses AI techniques for the purpose of natural language understanding (NLU) and generation (NLG), while NLP itself plays an important role as a component in collaborative and explainable AI. Advancing NLP is an AI challenge by itself. Key AI challenges facing the discipline today include (1) how to deal with the rich variation and cultural differences in language use and communication at the personal and group level in a data efficient manner; (2) how to optimise interactive language-based systems in extremely large, non-stationary state and action spaces; (3) how to achieve task and language independence, so that models developed for natural language understanding can be transferred across tasks and linguistic genres and languages with minimal re-training; (4) how to achieve naturalness in generated speech, responses, and narratives, using persona-based, emotional, and knowledge grounded content generation and understanding, (5) how to model and exploit explicit and implicit contexts for NLU and NLG, and (6) how to model concepts and meaning from large text and multimodal data.

4.6 INFORMATION RETRIEVAL

Information retrieval is concerned with connecting people to information. Search engines, recommender systems and digital assistants are prominent realisations of the discipline and of AI in practice. How can systems grasp people's information-seeking intent so as to be able to identify the right information for the right person, at the right time and in the right way? Key AI challenges facing the discipline today include (1) how to develop methods that learn to assess and improve their results through interactions with users; (2) how to develop reliable simulation environments that enable information retrieval systems to be reliably optimised using logged interaction data; (3) how to develop extremely data efficient learning methods that allow machines to identify and understand human tasks; and (4) how to develop mixed initiative retrieval methods that enable machines to understand when and how to hand over retrieval tasks to humans.



4.7 AUTONOMOUS AGENT SYSTEMS

The integration of capabilities and techniques from areas such as NLP, planning, search and vision is becoming the main research focus of the autonomous agents research field. The interaction between autonomous agents and the corresponding notion of social intelligence have become the focus of the multi-agent systems (MAS) research field. Important challenges for autonomous agents research are (1) to integrate AI techniques into a coherent decision-making architecture, and (2) to develop interaction and coordination models and techniques. Also, agents are often embedded into autonomous platforms, such as robots and cars; (3) truly autonomous systems require improvement in perception, manipulation, and navigation capabilities, as well as the development of sophisticated cognition and collaboration capabilities. The use of advanced machine learning techniques on autonomous platforms requires substantial experimentation, which often only becomes feasible through the use of simulations (e.g., for complex traffic situations in autonomous driving); (4) how can we assess to which degree these simulations are realistic and sufficient for learning or evaluation task at hand, and how can autonomous systems learn most effectively from such simulations?

4.8 AI SYSTEMS ATTUNED TO AND INSPIRED BY HUMAN COGNITION

To interact with humans in an effective, trustworthy and safe manner, AI systems must be aware of human cognitive abilities and mechanisms (see also cross-cutting

consideration III). At the same time, biological examples and human cognition have inspired the design of AI systems and formalisms, such as neural networks and rule-based reasoning systems. Human-computer interaction, biology and cognitive science are expected to continue to play an important role in the design of future AI systems. Research questions in this area include: (1) How can we distribute tasks between humans and automated systems, such that an optimal balance is obtained, taking the strengths and limitations of human perception and cognition into account? (2) How can we construct AI systems that dynamically collaborate well with humans over long periods of time, for instance by mimicking key aspects of human cognition? (3) How can we integrate human and machine learning effectively, e.g., in the context of personalised education, life-long and on-the-job learning, and compensation for cognitive decline in humans? (4) How can we effectively support a variety of stakeholders, including AI-experts, domain experts, decision-makers and people affected by the use of the AI system, each with their own requirements, strengths and limitations?

4.9 DATA DEPENDENCIES, QUALITY AND ENRICHMENT

Data-driven AI methods, such as machine learning techniques, rely heavily on the availability of data. In a way, dependencies on the data infrastructure are becoming more critical and more costly than code dependencies. It is important to understand these dependencies, although these are often difficult to analyse. How can we

establish trustworthiness of data and can AI itself be used to improve data quality, cleanliness, and timeliness? How can we decide which data, or which enriched data (features) to keep and which to discard, as they could be removed with no detriment? What data mechanism can be developed such that companies, institutions, consumers or citizens can view and control the correctness and quality of their data, and how their (personal) data is used? What are proper architectures for storing and processing of different (streaming, graph, ...) data amenable to AI processing algorithms? What infrastructure and machine learning algorithms are needed to develop and deploy AI in the case of distributed (personal) data? Methods and techniques from data science are important for the massive uptake of AI, including data analytics as the process that precedes a fitting AI system, and visualisation as the process that assists in the interaction with humans during design or operation of the AI system.

4.10 AI-DEDICATED HARDWARE

We are increasingly surrounded by hardware with embedded AI (smartphones, autonomous vehicles, smart prosthetics), requiring an optimal balance between software and hardware. Successful autonomous behaviour depends on a reliable continuous feedback loop between sensors and activators. AI applications embedded in a human setting (at home, at work, in healthcare) must be well-aligned with human needs and limitations. Many of today's successful AI implementations require large datasets at the cost of huge energy consumption. Hence research is needed into alternative hardware architectures dedicated for AI, such as in parallel computing, dynamical systems, neuromorphic engineering, and quantum computing. Can AI-dedicated hardware be designed that successfully operates with sparse data and at low energy levels? Can AI implementations benefit from parallel, analogue hardware combined with or in contrast with serial, digital chips? Can hardware with low-level uncertainty, ambiguity and contradiction lead to reliable behaviour at the high level? How can artificial hardware be inspired by natural systems? Can vulnerabilities and attacks be addressed by coordinated software hardware designs?

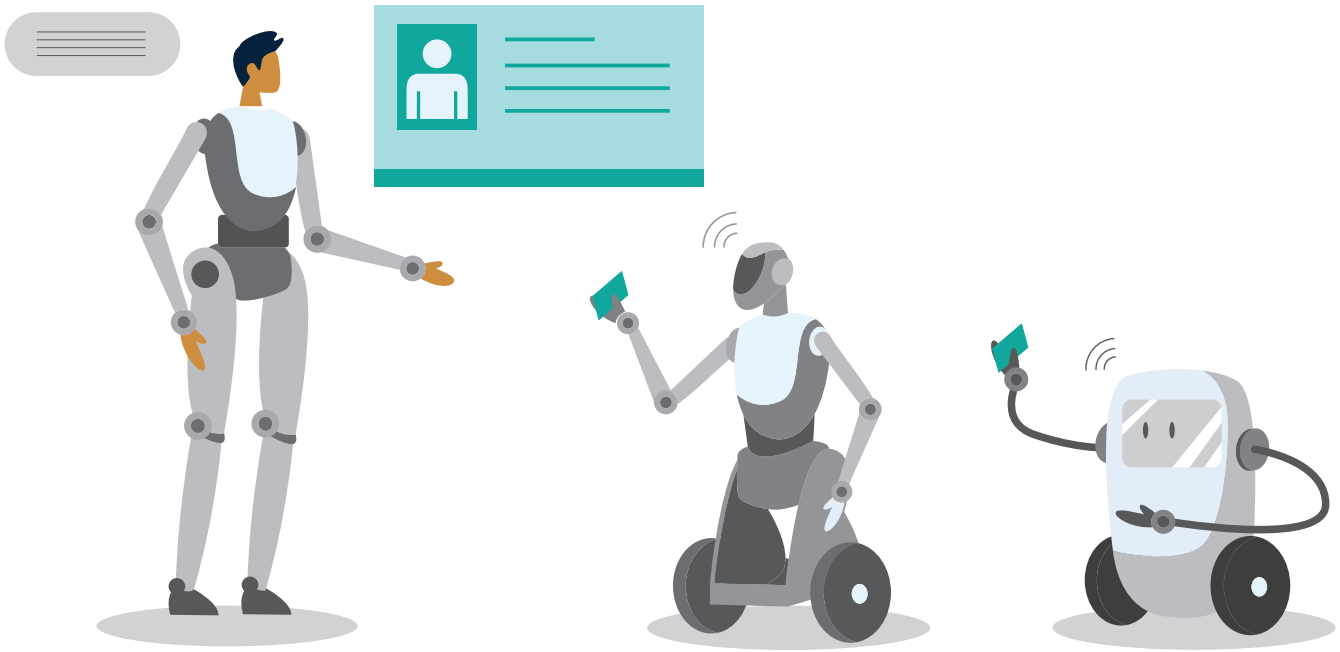
4.11 ETHICAL DIMENSIONS OF AI

One of the most profound implications of AI for society is that decision-making becomes an automated process, distributed between humans and machines, which is often opaque or lacking in transparency. There is a clear need for optimising and understanding these processes in view of their widespread societal impacts. Given the potential for negative consequences alongside the positive benefits, it is important to consider what principles determine fairness in AI applications such as automated decision-making (ADM)? How does ADM translate into technological design, legal rules and individual or societal effects? A key aspect of elucidating the

essence of what fairness in AI means is identifying patterns, commonalities and distinct features across different sectors and contexts across three dimensions: (1) Procedural dimension. Which conditions may safeguard fairness? (for instance, the right to a human in the loop, mechanisms for discovering bias and solutions for the lack of accuracy, contestability and meaningful explanation). (2) Inter-relational dimension. What is the optimal division of tasks between human and automated decision-makers to achieve optimal trust and individual and societal acceptability? (3) Ethical-legal dimension. What should be allowed and what should be banned as societally (un)acceptable forms of AI and ADM?

4.12 LEGAL REQUIREMENTS FOR AI

The existing legal order (including the fundamental rights framework, rules on data protection, non-discrimination, intellectual property, protection of business secrets, competition law, legal liability and due process) are important boundary conditions for AI development, implementation, management and maintenance. These frameworks must be geared towards optimising benefits and addressing harms from AI and ensuring responsible (use of) AI. Defining legal requirements must be informed by a solid understanding of relevant AI technology, and the way technology, humans and the law interact. To what extent are the existing frameworks



adequate, and where are new regulations needed? How can effective control and governance of AI systems be realised, including the allocation of responsibility, duties of care, procedural fairness and actionable rights? How can we find the proper balance between legal abstraction and detailed instructions that can inform technology design? How do we deal with (proprietary) AI technology and data as sources of considerable economic and political power to promote independent research, innovation, fair competition, choice and the realisation of fundamental rights?

4.13 SOCIETAL CONTEXT OF AI

The implementation of AI algorithms and the use of data are intrinsically rooted in socioeconomic and technical-legal choices and have consequences at the level of individuals and groups. Accordingly, ethical, legal and social implications can never be an afterthought, but should be accounted for right from the start. Because of this inter-relatedness, AI should be framed as a socio-technical system; a product of social systems that will at the same time impact those social systems. This also requires us to make visible the potential choices and influences that sometimes reflect conflicting norms and values (e.g. fairness, privacy, autonomy, dignity, accountability, due process, etc.). Importantly, technological developers, sectoral professionals and users are collectively responsible for thinking through the impact of AI on society and for ensuring and inclusive process through a diverse set of opinions. This requires understanding the way humans, machines and society interact and jointly behave. How can human-centred AI be integrated to strengthen societal organisation, institutional practices and democratic processes? How can we involve professionals from specific societal sectors in the articulation of conditions for developing AI and involve them in co-design of algorithmic decision-making? How can we identify the conditions that enable AI applications to make a change for the better and avoid unintended consequences and new individual or systemic vulnerabilities?

5. IMPACT ON APPLICATION DOMAINS

Solutions to the grand research challenges and cross-cutting considerations will serve the AI needs of many application domains. In the context of this research agenda, we have selected several representative examples. Given the enormous potential of AI in many different application domains, this overview does not have the pretention to be complete. The examples we selected fit well with the Dutch, European and UN Sustainable Development Goal challenges, with societal and industrial impact, with scientific discoveries and with an emphasis on the common good. We also note that for the uptake, implementation, and impact of methods in a particular application domain, AI researchers will actively work together with domain experts (scientists/practitioners) from the field.

5.1 HEALTH AND WELL-BEING

AI can support early detection and preventive strategies at the population level by uncovering correlations and patterns between health data, health status and disease onset. Citizens can be empowered to take a more proactive approach to their health, by being able to monitor their health status, disease risk, and the effectiveness of lifestyle changes. AI will augment clinicians with tools to aid in the diagnosis and clinical decision-making, leading to more precise, personalised treatments. It will also improve workflows in healthcare institutions, by providing the right information at the right time. AI systems need to be able to deal with heterogeneous knowledge, data of various quality, consider the end user needs and perspective, and explain why certain diagnoses and health recommendations are made. AI can also complement human and domain medical knowledge in studying the complex interrelation between genetic liability, lifestyle factors, environmental data, and health. Therefore, in this field there is a particular need for AI that combines data-driven approaches with prior domain knowledge, and a health data infrastructure in which all these different distributed data sources can be accessed (FAIR data and distributed learning). In addition, AI solutions will need to be able to adapt to the constant changes in medical technology and practice, which requires methods for continuous learning (learning healthcare system).

5.2 SAFETY AND SECURITY

For critically assessing security and safety situations, professionals are dependent on correctly analysed sensor and internet data.

In this way they can respond adequately to threats and hazards in industrial context, public safety, cybersecurity and national defence. The available data is heterogeneous and varies from camera imagery to chemical sensor readings, and from hate speech to DDOS attacks on internet services. A specific challenge in this area is the imbalance of the data: most of the collected data is innocuous, just a tiny fraction represents interesting case data due to malicious behaviour or random accidents. AI algorithms learn poorly from such unbalanced classes, and the combination with domain-specific knowledge is a necessity. Not only the reliability of the data and the verification of decisions are important, also the timeliness plays an important role in sudden situations. This represents an even bigger challenge for first responders' interaction with AI systems and explanation of the “right” action that needs to be taken. In view of the legal framework in this domain, the combination of automated reasoning and machine learning is going to be key to progress. The sharing of (confidential) data for use in AI and the responsibility for actions taken have a high priority, especially with the trend towards autonomous platforms such as drones in the area of safety and security.

5.3 MOBILITY AND TRANSPORT

In the Knowledge and Innovation Agenda for the Top Sector Logistics, a separate agenda has been put forward to address mobility issues and to make them future-proof. It is believed that the technological developments in the areas of digital and autonomous technologies will have a huge impact on increasing the efficiency, safety and sustainability of mobility and transport as well as decreasing its costs. Furthermore, smart shipping and autonomous shipping are key areas of interest identified. Decision-support is also considered to be an important building block in the realisation. The agenda states that many components are already available but need to be integrated. At the same time, substantial challenges - particularly in the areas of robustness - remain to be addressed by fundamental research. Progress in this area will require demonstration projects that integrate existing and new AI technologies. Research and engineering need to focus on some fundamental developments in key technologies considering all TRL levels. In this application domain, embedding of AI in sensor technology and robotic platforms play an important role too.

5.4 AGRI-FOOD

AI can be used to improve our understanding of biodiversity and the effects of climate change, for automating the control of greenhouses, for monitoring the well-being of livestock, optimising energy and usage of other resources in this domain (e.g., water), and for providing support for healthy and sustainable food choices. For optimising resource usage, AI techniques can be used to automatically learn from data collected by monitoring systems to improve sustainability in the long term. The combination of remote sensing, robotics, big data and AI has great

potential in this regard. The Dutch agri-food sector is well-known for its innovative potential. The development process of new market products in this sector can also benefit from combining its expertise with AI to enhance its innovative potential. Moreover, AI can provide techniques for optimising human decision-making and enhancing collaboration between existing and future intelligent systems by means of hybrid AI, i.e. by combining human and machine expertise in the production process (for instance, decision-support systems, data visualisation). The application of AI can also be leveraged for improving and personalising consumer lifestyle support and decision-support for sustainable food consumption.

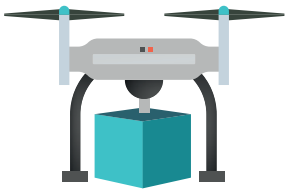
5.5 AI FOR COMMON GOOD: SUSTAINABLE, ENVIRONMENTAL, CIRCULAR

In 2015, the UN General Assembly adopted the Sustainable Development Goals (SDGs) as part of an international framework to foster sustainability of environment and society. This framework consists of 17 SDGs and 169 targets to be achieved by 2030. Among them are goals such as "Ensure sustainable consumption and production patterns" (Goal 12), and "Ensure access to affordable, reliable, sustainable and modern energy for all" (Goal 6). While AI by itself is certainly no silver bullet capable of solving all SDGs, AI techniques, in combination with other technologies (such as remote sensing), are expected to play a key role in achieving the SDGs, by critically enabling progress in: precision agriculture, monitoring of environmental indicators locally, regionally and globally, tracking forest density, minimising food and energy wastage, and increasing energy and resource efficiency. AI is also being used to predict climate-related disasters and in preventative healthcare programmes. Of equal importance is the need for AI to be developed and deployed using sustainable and environmentally friendly methods (e.g., sustainable computing, next-generation batteries, solar fuel to better harness energy from natural resources), so as to ensure the utility of AI is not undermined by its environmental impact.

5.6 SERVICE INDUSTRY

More than any other industry, the consumer-facing service industry (accommodation, finance, health, lifestyle, tourism, retail, transportation, etc.) is subject to sudden major disruptions due to AI. With massive volumes of transactional data being generated as a natural side effect of operational services, there is tremendous potential for innovation through AI. Users' long-term interests can be mined from logged transactional data, which can inform highly personalised recommendations, increasingly communicated through (AI-powered) digital assistants. Through suitable levels of online exploration, modern AI algorithms are increasingly well equipped to detect users' short-term interests and infer changes in tastes and preferences. AI promises to enable the service industry to tailor offers to individual consumers with high levels of precision. Significant

challenges remain, however, challenges that require substantial foundational advances in AI: in counterfactual evaluation and learning from historical interactions, in multimodal interactions, in automatically generating domain-specific knowledge graphs so as to support conversational agents, and in using AI for real-time forecasting of individual and population-level preferences.



5.7 SMART INDUSTRY

Digitisation is a dominant development in the manufacturing and supply chain industry. Robotisation, big data and artificial intelligence are radical innovations, aimed at increasing the competitiveness, servitisation, and new products and services being brought onto the market. A success case is the use of AI for predictive maintenance, in which data collected by (multiple) remote-monitoring sensors is used to monitor the performance and condition of equipment under normal operation so as to reduce the likelihood of failures. With artificial intelligence algorithms more and more industrial and logistic processes can be performed automatically. The application of machine learning makes it increasingly difficult for people to understand what is happening inside a machine. Specific knowledge in companies about process and machine peculiarities, optimisation, and interdependencies must be included in AI solutions and in the interaction between employee and AI. The shifting balance between tasks to be performed by humans and AI systems and the resulting impact on responsibilities and the future of work are major challenges for the development of smart industry.

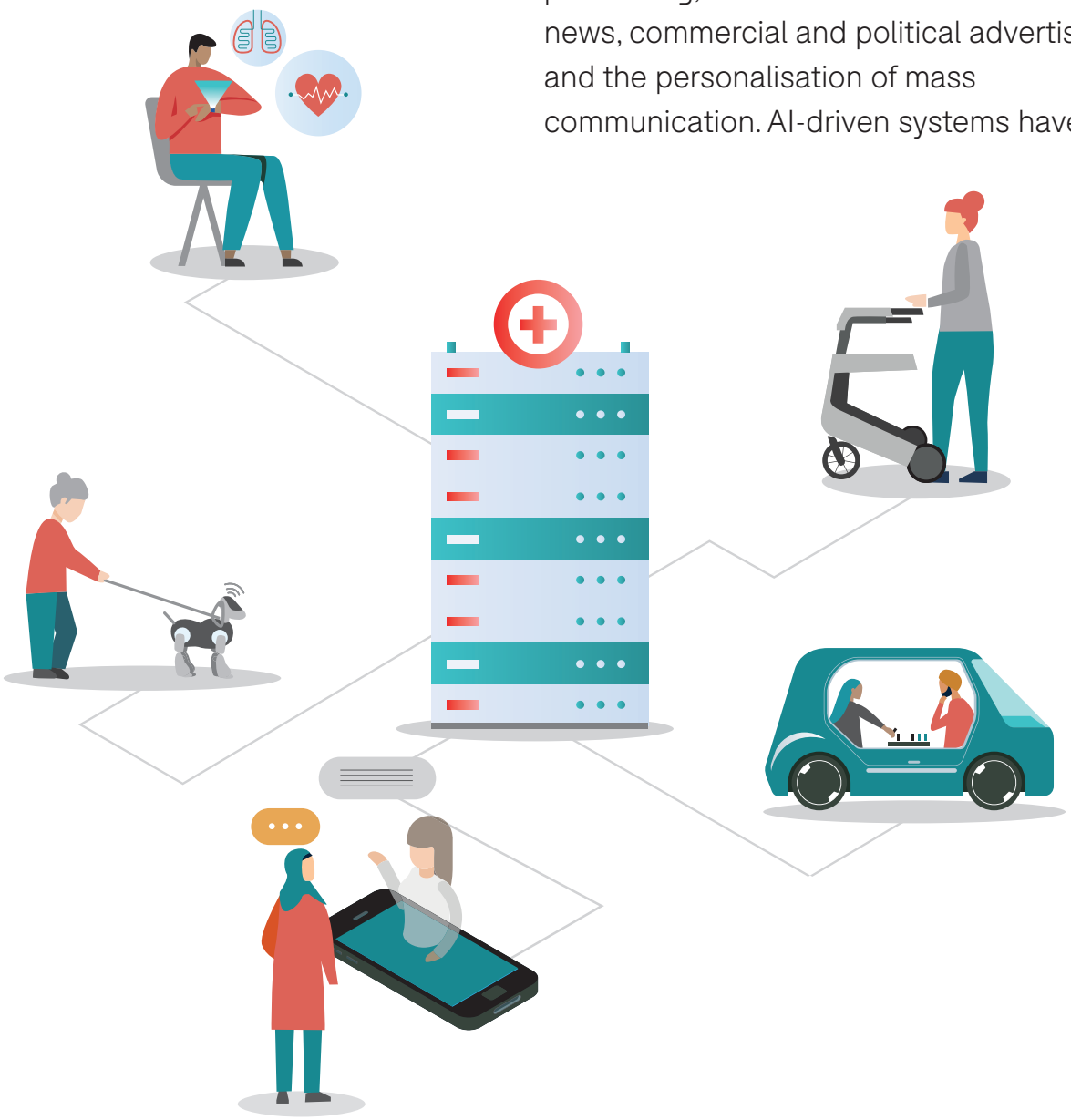
5.8 LEGAL DECISION MAKING

NLP-based legal text analytics and information retrieval techniques hold great prospect for law. However, to fully utilise their potential, these techniques must be combined with normative legal argumentation and adequate forms of structuring legal data to support humans in making sense of the text analytics results. Legal actors (from courts to law firms and

government bodies) increasingly use search engines and machine learning algorithms for predicting outcomes of cases, risk profiling, monitoring citizens and making decisions about them. While these uses of AI can have many benefits, they also raise significant concerns with regards to fundamental rights, procedural and substantive justice and the democratic division of powers through checks and balances. The data on which these decisions are based may, and often do contain biases that are both unwanted and unknown. Contributing to these concerns is the non-transparent, black-box nature of some AI reasoning and decision-making, and the lack of adequate frameworks to hold algorithmic decision-making power to account. Explainability and responsibility are critical success factors so that AI decisions can be understood, and critically examined and, where necessary, challenged and adapted.

5.9 MEDIA AND DEMOCRACY

AI challenges for media and democracy revolve around the automated production, processing, moderation and distribution of news, commercial and political advertising, and the personalisation of mass communication. AI-driven systems have the



potential to fundamentally change the media as we know it, and forms of automated text generation, moderation or deep fake fabrications could seriously impact the information order, politics and democracy at large. The growing reliance on automated content moderation as a response to misinformation and unlawful content raises serious fundamental rights concerns. Elections are increasingly informed, but also manipulated with the help of AI including automated content generation, advanced data analytics and the use of data-driven recommendations. The relationship between the news media and audiences is also shifting in fundamental ways, and the power to shape, or manipulate, individuals' news exposure brings new opportunities, but also responsibilities and challenges for professionals. On the other hand, we see that AI is used to handle abundance of large media streams, detect disinformation and misinformation and capture dynamics and spread of information in online debates and filter bubbles. The impact of AI on media and democracy demonstrates the importance of security, reliability, explainability and trustworthiness of AI solutions, and the need for a solid legal framework to guide their use.

5.10 NEXT-GENERATION SCIENTIFIC DISCOVERY AND ENGINEERING

Across all sciences and engineering disciplines, computational methods are starting to play a key role not only in the analysis of data and the empirical testing of hypotheses, but also in the generation of models, hypotheses and new artefacts. AI techniques are beginning to have a major impact on the way scientific knowledge is produced, tested, refined and revised. Prominent examples for this can be found in many areas, ranging from astronomy to evolutionary biology, from materials science to particle physics, from climate science to drug design, and from the design of smart products to the engineering of advanced high-tech systems. In these and many other disciplines, techniques from various areas of AI, including machine learning and pattern discovery, automated reasoning, planning and search, but also robotics, computer vision and natural language processing, will have a transformative impact on the way scientific studies are conceived and conducted, thus enabling scientific discoveries that would not have been possible otherwise. Due to the momentous impact of scientific discovery on a broad range of engineering disciplines and application domains, we expect the use of AI in this context to be of large and long-ranging economic and societal importance. We also see next-generation scientific discovery and engineering as particularly well-suited application area for human-centred AI techniques and systems, since the combination of human and machine intelligence is going to be the driving force of scientific and engineering progress for the foreseeable future.

6. RELATION TO OTHER AI AGENDAS

The grand AI research challenges and cross-cutting considerations formulated in this agenda by the Dutch AI research community, are well aligned and consistent with the strategic development directions of national and European AI agendas.

The Dutch government's AI strategy (SAPAI)[2], and the national AI coalition (NLAIC)[1,3] emphasise that AI benefits for Dutch society and economic sectors will only happen if we strengthen our position in research and innovation, and if the Netherlands increases its efforts in attracting and retaining talent. Similarly, the top sector[6] multi-annual programme (MJP) entitled "National AI research centre" outlines the need for a dedicated research programme bringing together the on-going successful collaborations such as the ICAI network[7], CLAIRE and ELLIS, and strengthening these initiatives through a large fundamental AI research programme. This AIREA-NL agenda provides details about the challenges that can and must be addressed by the Dutch academic community.

The recently awarded NWO Gravitation programme proposal Hybrid Intelligence[8] focuses on aspects of grand challenge 3.3: AI Systems and Humans. It tackles the problem of how to organise collaborative teams of human and artificial agents to solve complex tasks in science, health and education. Through its RQ-3.3, AIREA-NL is well aligned with the Hybrid Intelligence agenda.

AI plays an important role as an enabling technology for the prioritised societal and departmental challenges in energy and climate, food and agriculture, healthcare and well-being, and safety and security. Many of the forthcoming multi-annual, mission-oriented programmes (MMIPs) refer to the use of current and existing AI solutions, and the need for solutions for cross-cutting considerations such as explainability and responsibility. None of these programmes addresses fundamental challenges in AI, even though solutions for these will be needed in tomorrow's AI applications. In that sense, AIREA-NL fills an important gap in the national AI investment landscape by focusing on the fundamental research needed for successful and responsible applications of AI, today, tomorrow, and thereafter.

6. https://www.topsectoren.nl/innovatie
7. https://icai.ai/
8. https://www.nwo.nl/en/research-and-results/programmes/nwo/gravitation/awards-2018-2019.html

In the European context, several agendas and strategies have been proposed. Many, if not all of these agendas embrace the urgency of strategic investments in fundamental AI technology, industrial and societal applications of AI, as well as the importance of a robust regulatory framework. With the Commission's Communication on the "European Initiative on AI" and on the "Coordinated Plan on AI", and with the report of the independent High Level Expert Group on AI[9], Europe has made it abundantly clear that investing in AI, a solid regulatory framework and guaranteeing that AI systems are in full respect of fundamental rights and core European values, must be a key political priority. The European Commission warns that the growing dependency on technologies developed outside Europe is a risk for our economy, society and democracy. Digital leadership and removing the European dependency on technologies developed outside Europe is a guiding objective for the new Commission. This perspective strongly reinforces the urgency for strategic Dutch investments in the grand AI research challenge outlined in the AIREA-NL agenda.

CLAIRE is a large, pan-European organisation that aims to ensure European excellence in human-centred AI, leveraging strength across all areas of AI across all of

9. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Europe, and providing guidance to the European Commission and national governments in their investments in AI. CLAIRE emphasises the importance of a broad range of AI techniques for solving challenges from diverse applications of AI. ELLIS is a European network focussed on building and maintaining European leadership in machine learning. Both organisations have a focus on strengthening basic, interest-driven research as a foundation for achieving economic impact. CLAIRE and ELLIS are both strongly represented in the Netherlands.

There is a vibrant culture of not-for-profit foundations sharing strong links between academia, policy makers and the general public, such as ALLAI (the Dutch Alliance on AI), UNICRI (United Nations Interregional Crime and Justice Research Institute) Centre for AI and Robotics and FRR (Foundation for Responsible Robotics). Each of these organisations concentrates on various societal challenges facing AI (e.g., security and crime), inclusion of the public in discussions about AI, and protection of human rights in the design and implementation of AI.

AIREA-NL is well aligned with the visions pursued by CLAIRE, ELLIS and national foundations, and will benefit from the activities of these organisations in the Netherlands.

The German AI agenda[10] revolves around the responsible development and use of AI to serve the good of society. It aims to

10. https://www.ki-strategie-deutschland.de/home.html?file=files/downloads/Nationale_KI-Strategie_engl.pdf

integrate AI in society in ethical, legal, cultural and institutional terms in the context of a broad societal dialogue and active political measures. Although the German agenda does not outline specific research challenges, it will further develop existing Centres of Excellence for AI, and it will create at least 100 additional professorships for AI to ensure that AI has a strong foothold within the higher education system.

The French AI strategy[11] builds on five pillars, from building a data-focussed economic policy to ethical considerations of AI. It includes a strong pillar that aims to build an agile and enabling research programme around a network of independent but coordinated interdisciplinary AI institutes. These institutes focus on specific aspects of AI, and have a strong focus on an interdisciplinary approach, notably by including social scientists.

The Flemish AI agenda[12] is organised around three pillars: (1) fundamental research into technology across the full spectrum of all areas of AI research, (2) industrial applications, and (3) education, raising awareness, and ethics. Plans to recruit 400 PhD students are being rolled out during 2019.

The United Kingdom was one of the first European countries to adopt a full-fledged AI strategy[13]. The AI Sector Deal is organised in five pillars: Ideas (for research and development), People (for talent development, including 16 centres for doctoral training, delivering 1000 new PhD students over the next five years), Infrastructure (for digital and data infrastructure), Business environment (for AI business development), and Places (to empower local AI ecosystems around the country).

Finally, AIREA-NL aligns well with four of the five main enablers of the Strategic Research, Innovation and Deployment Agenda for a European AI PPP[14], namely "continuous and integrated knowledge", "trustworthy hybrid decision-making", "physical and human action and interaction", and "system, methodology and hardware". In conclusion, it is the AI scientific community's hope and aim to have the research-related aspects of the forthcoming Dutch national AI strategy closely aligned with AIREA-NL. That will enable the Netherlands to position itself among those countries and initiatives that aim to be leaders in AI technology, in strengthening their economy thanks to AI, and in the seamless integration of AI developments in their sociocultural fabric.

11. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf
12. https://www.ewi-vlaanderen.be/nieuws/30-miljoen-euro-voor-vlaams-actieplan-artificiele-intelligentie
13. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal
14. https://www.eu-robotics.net/cms/upload/downloads/ppp-documents/AI_PPP_SRIDA-Consultation_Version-June_2019_-_Online_V1.3.pdf

Dutch Research Council (NWO)

NWO Den Haag
Laan van Nieuw Oost-Indië 300
2593 CE The Hague

+31 (0)70 344 06 40

NWO Utrecht
Winthontlaan 2
3526 KV Utrecht

+31 (0)30 600 12 11

www.nwo.nl