Controllable Artificial Intelligence
A Human-centered Approach

Mehdi Dastani
Utrecht University

30 May 2022

Mevrouw de Rector Magnificus,

It may be natural to begin this lecture with a reference to Charles Babbage, John von Neumann, or Alan Turing, founders of computer science. Or John McCarthy who coined the term artificial intelligence at the Dartmouth summer school organized in 1956. If I want to emphasize the origin of philosophical issues related to artificial intelligence, I can even go back to the ancient Greeks. However, I would like to go back in time a little further, to the creation story, not because I believe the story is true, but because it reflects an ancient and longstanding human desire. In the earthly paradise, there was no need to work. The punishment meted out to Eve and Adam after the so-called 'Fall of Man', was this: "By the sweat of your face you will earn your daily bread henceforth".

The reason I am citing this is to emphasize how much 'work', or maybe I should say 'some type of work', has been regarded as a punishment since ancient cultures. The rest of the history is largely about human struggle to reduce its punishment by recreating a paradise on earth. The idea of constructing an artificial paradise by either exploiting others to do the work, whenever there is a way around legal, social and ethical concerns, or by mechanizing and eventually automating it, has always been the utopia of humankind.

We are now at the edge of the age of artificial intelligence, which is rapidly and radically transforming the world around us and is redefining the role of human beings. The development of AI has accelerated due to the massive digitalization and datafication of our practices, coupled with the power of computing capabilities. Thanks to these developments, we are now witnessing alternative ways of living, with services and tools that enrich and improve our daily activities. The ongoing transitions towards digitalization, datafication, automation, and autonomization, the core of artificial intelligence in my opinion, will not be smooth, however. Without doubt, these transitions will have a significant impact on almost all societal sectors, from health and education, to mobility and critical infrastructures. It is therefore imperative, that we address some key questions, such as: how will these transitions change our economy, the future of work, our

history, and our culture? What impact will they have on environmental issues, and how will they affect human autonomy?

Fortunately, these and other critical questions are receiving more and more attention, not only from the AI community, but also from other scientific disciplines, governments, civil society, and the public. There is now a growing awareness that some of these questions are not aimed at artificial intelligence as such, but at digitalization, datafication, and automation. This awareness is reflected, for example, in a recent report from the Netherlands Scientific Council for Government Policy [46]. They advise the Dutch government on how to embed AI in Dutch society. In their report entitled "Mission AI: The New System Technology", they identify what needs to be done to embed AI as system technology, and they make recommendations on how to do this. One of their recommendations is to demystify AI, i.e. make it clear and explicit, particularly to the general public, what artificial intelligence actually is. Demystification of AI should tackle the overly optimistic and pessimistic images of AI, characterise AI properly and distinguish it from digitalization, datafication, and automation. Finally, it should learn to focus on the right questions and face the key challenges directed at AI.

A distinguishing characteristic of AI systems, and a main source of concern, is their autonomy; that machines perform tasks that involve decision-making on behalf of human decision-makers. Autonomous systems can be supportive, but they can also be experienced as limiting human autonomy, or even destructive. Think, for example, of amazon services that decide for you which products you should consider to buy, when you visit their website. Or self-driving cars whose decisions may cause severe accidents. It is therefore essential, that humans face the challenge of steering and controlling AI systems to improve their lives, rather than letting themselves be controlled by these systems.

My lecture today will focus on human-centered approach to artificial intelligence and how this approach can keep AI systems under control. But what is human-centered approach to AI? And what does it mean to keep AI systems under control? Obviously, the answers depend on who you ask. One may use 'human-centered' to focus on the ethical, legal and social impacts of the AI technology. They may advocate regulations and laws to retain control over the technology. Others may use 'human-centered' to emphasize the virtues such as understandability, explanability and traceability that, when possessed by the AI systems, make them controllable. They advocate the use of specific techniques, methods and approaches in AI that support these virtues. In this lecture, I will share with you my thoughts on the second view. Please allow me to reiterate some episodes in the history of AI that are important for the rest of this lecture.

1 Some episodes in artificial intelligence history

Back in 1950, Alan Turing posed in his seminal article 'Computing Machinery and Intelligence' [57], the fundamental question 'Can machines think'? He recognized this as a challenging question, and argued that for a proper answer, one needs to come up with definitions of the terms 'machine' and 'think'. For the term 'machine', he had already proposed a precise mathematical definition, which to this date is the very foundation of computing machinery. However, for the meaning of the term 'think', or 'Intelligence' in general, he had no precise mathematical definition. Probably, influenced by his contemporary philosopher Wittgenstein, Turing argued that any definition of the term 'intelligence' should be as close as possible to the 'normal use' of this term. Interestingly, he argued in his article that defining the meaning of the words by analyzing their uses is dangerous, as the answer to the question 'Can machines think?' should be sought in a statistical survey, which he thought to be absurd. Probably because at that time there was no proper survey, machine learning techniques, and enough computing power.

Instead, he became inspired by the Victorian-era imitation game, where one tries to identify the gender of another person by asking questions. Turing proposed a similar game, nowadays known as the 'Turing test', where a human interrogator who is physically separated from another human being on the one hand and the supposedly intelligent thinking machine on the other hand. The human interrogator can ask directed questions to determine which of the two is the human being and which one is the machine. If the interrogator is unable to correctly identify the machine, then the machine can qualify as intelligent. Turing's qualification of intelligence can be best presented by the following quote:

> "I believe that in about fifty years' time, it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well, that an **average interrogator** will not have more than **70 per cent**, **chance** of making the right identification **after five minutes of questioning**." *(the emphasis is mine to highlight the empirical/descriptive qualification of intelligence)*

Seventy years later, there is an established scientific discipline called Artificial Intelligence. After all these years, and despite the intense period of activities, developments, and debates, there are still different views of what AI is and how AI systems can be characterized. For example, some regard AI as a science aimed at understanding, modelling and enhancing intelligence, that is considered as a natural phenomenon. Others explain AI more pragmatically as the study of the computational modelling of increasingly complex tasks and functions for which human intervention is required. An extreme example of this pragmatic view was formulated as Larry Tesler's Theorem that describes artificial intelligence as whatever that has not been done yet [31]. A more recent description, which is also known as the agent view [48], considers Artificial Intelligence as referring

to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. This definition has also recently been used in official reports such as that of The High-Level Expert group of the European Commission [30] and the Netherlands Scientific Council for Government Policy [46].

These and other descriptions of AI are based on specific, often unarticulated, assumptions about 'what counts as a (scientific) AI model?', 'should an AI model provide insights about the modelled phenomenon in terms of underlying principles, or is it sufficient to simply simulate that?', 'when would a principle be understood and accepted as explanatory and insightful?', 'what is the difference between AI as science and AI as engineering?', and 'could an AI engineering success be considered as proof that the engineered phenomenon is correctly understood and modelled?'. These issues have been thoroughly discussed in numerous debates throughout the relatively short history of artificial intelligence. For example, the Chinese room argument formulated by John Searle in 1980 [50], which was a reaction to the Turing test, questioned the plausibility of simulating models. Searle asked whether it is acceptable to attribute intelligence to a machine that generates based on an input/output table, and I quote:

> "Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else ... A system, me, for example, would not acquire an understanding of Chinese just by going through the steps of a computer program that simulated the behavior of a Chinese speaker." [49] (p.17)

The following philosophical debate on 'Strong' versus 'Weak' AI, has been conducted to the very extremes, by questioning whether a machine can have a mind and consciousness. A typical philosopher in this regard is Hubert Dreyfus who put forward in his book "What Computers still cannot do" a phenomenological argument against the possibility of machines having a mind and consciousness. He opposed rationalists such as Descartes and Leibniz who thought of the human mind as a mechanical device with the capacity to form representations of all domains of activities. Dreyfus argued, and I quote:

> "... intelligence requires understanding, and understanding requires giving a computer the background of common sense that adult human beings have by virtue of having bodies, interacting skillfully with the material world, and being trained into a culture." [23] (p. 3)

According to Dreyfus, common sense is a form of knowledge that determines what is relevant and important at each moment in time. Common-sense knowledge, Dreyfus argued, is closely linked to human biological and psychological conditions and concerns. Dreyfus concluded that it is improbable that a device

can be built that is sufficiently like us that it can learn the same common-sense knowledge as us and act in our world in the way we do.

Since I agree with most of the Dreyfus' arguments, I will stick with the weak AI position, and forget the mind and consciousness questions altogether. I rather focus on the debate about modelling intelligent behavior [39]. More specifically, the question of whether a model, that simulates intelligent behavior, is satisfactory, or should the model be based on underlying explanatory principles. One example is the debate that took place at the symposium 'Brains, Minds, and Machines', held during MIT's 150th birthday party in 2011.

"Technology Review" reports on this symposium, and claims that Chomsky has mocked researchers in machine learning who use purely statistical methods to produce behavior; they would mimic something in the world without trying to understand it. Chomsky seemingly takes the anti-simulating stance by stating, and I quote:

> "It's true there's been a lot of work on trying to apply statistical models to various linguistic problems. I think there have been some successes, but a lot of failures. There is a notion of success ... which I think is novel in the history of science. It interprets success as approximating unanalyzed data." [44]

According to Peter Norvig, Chomsky believes that statistical models of natural language should be seen as an engineering success that is not relevant to science. And that collecting and approximating linguistic facts by statistical models does not provide insights to the underlying principles that matter in science. In this discussion, Chomsky seems to categorize statistical models as mathematical tools that can approximate the input/output table in Searle's thought experiment.

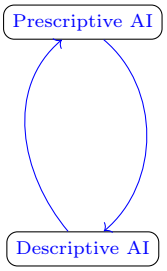
The debate on the nature and aims of AI is now shifted to technical discussions on the potentials and challenges of various approaches and methodologies in AI that are classified by dichotomies between model-driven versus data-driven AI, AI through reasoning versus AI through learning, and symbolic versus subsymbolic AI. Although these dichotomies seem useful for bringing structure in the AI literature, the practice of the AI research, as I will argue, is less sharp, with lots of overlap and mixes.

2 Dichotomies in Artificial Intelligence

Despite subtle differences, the dichotomy between model-driven versus data-driven AI, or to some extent 'AI through reasoning' versus 'AI through learning', can be seen as a reincarnation of the general dichotomy between prescriptive and descriptive approaches in science, now applied to the field of artificial intelligence.

Prescriptive AI tends to establish facts, laws and claims that prescribe how intelligent behaviour ought to be. They prescribe moral, economic, epistemic, or rational rules that intelligent behaviour ought to respect. For example, decisions should be economic rational, outcomes should be fair, knowledge should be consistent, and new observations should lead to minimal information revision. Prescriptive models are often evaluated in terms of their theoretical adequacy, i.e., the degree to which they provide acceptable idealizations and are aligned with acceptable norms.

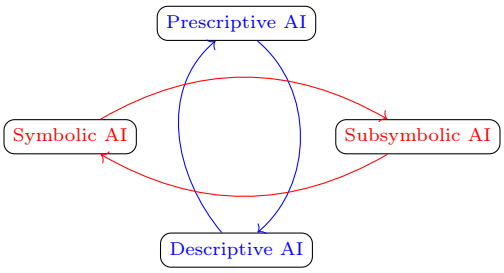
On the other hand, descriptive AI tends to be factual, data-driven and experimental. They establish facts, laws and claims that describe intelligent behaviours as manifested in practice. Intelligent behaviours are described as patterns that occur and reoccur in practice. They are modelled by learning from the corresponding data. Descriptive AI-models are often evaluated in terms of their empirical validity as measured by various metrics such as precision, recall, accuracy and entropy, and of course their applicability.



As also argued by Bell, Raiffa, and Tversky in their book "Decision Making: descriptive, normative, and prescriptive interactions" [47], the distinction between prescriptive and descriptive models is not always as sharp as one would expect. Prescriptive models are often used as first-cut descriptive models. They can go through successive modifications to make them useful for descriptive purposes, for example, for prediction. Applying this view to decision-making in AI, one can argue that prescriptive rational decision models can become more useful/applicable for descriptive purposes, when calibrated to some relevant data, by for example using machine learning and optimization techniques. As I will explain later in this lecture, we have applied such an approach in an agent-based simulation project, where agents' decisions, modelled by rational decision rules, are calibrated to the behavioral data using optimization techniques.

On the other hand, descriptive models can be modified so that they become closer to what some analyst might believe are proper norms for wise, optimal and rational behaviour. For example, a descriptive model built by a clustering method using the proximity of embedded elements can be refined using explicit domain knowledge and rational axioms. We have applied such an approach in the Golden Agents project, which I will report later in this lecture.

There is yet another fundamental dichotomy in AI: symbolic versus subsymbolic AI. The correspondence between compositional syntax and semantics, the fact that ideas can be represented by symbols that can be manipulated through meaning-preserving calculi, is central to the symbolic AI. This differs from subsymbolic AI, where the correspondence between input and output, and thus the system's functionality and performance, is central.



The distinction between symbolic and sub-symbolic AI is also less sharp than one would expect. This is evidenced by, for example, embedding techniques, used nowadays in machine learning and neural networks, where each vector represents a real-world entity. But the fact that vectors can denote entities is not enough to say that embedding systems are symbolic. After all, it is unclear how embedding systems can ensure compositionality in the way symbolic systems do.

Despite the lack of a sharp distinction, a question is how symbolic and subsymbolic techniques can be integrated to complement and strengthen each other. Some AI researchers argue that symbolic techniques are more suitable for modelling high-level cognitive functions, such as reasoning and planning, while subsymbolic systems are more suitable for modelling low-level perceptual tasks, such as image recognition. This may be true, but the integration of symbolic and subsymbolic AI has more potential that goes beyond this coarse functional characterization. For example, faulty or imprecise symbolic knowledge can be embedded in metric spaces where the distorted similarity between domain elements can be examined using sub-symbolic clustering techniques. Conversely, one can start with embedded knowledge and use symbolic knowledge to improve its quality. I will give some examples later in this lecture.

So far, the techniques, methods, and approaches in AI. But, what is now human-centered AI and how does it relate to these distinctions?

3 Human-centered Artificial Intelligence

Recent developments and applications of data-driven methods, learning-based techniques, and sub-symbolic black-box approaches have made many to hope or

fear the future impact of AI. There is now a growing community that calls for human-centered AI.

Within this community, some advocate the regulation of AI technology, even through law, to control its ethical, legal, and social impacts. They may formulate guidelines and assessment procedures that should be followed when designing, developing, and deploying AI systems. They would ideally enforce these guidelines and assessment procedures by law. An example of this is the Human Rights and Algorithms Impact Assessment instrument [28] that is being developed by Prof. Janneke Gerards from the Utrecht University School of Law and her colleagues from the Utrecht Data school. This instrument is intended to enable the design, development and deployment of algorithms that make well-informed and responsible decisions, sometimes on behalf of human stakeholders. Others in the human-centered AI community focus on the collaboration between human and AI systems. They advocate high levels of understanding and control over AI systems. This view is best articulated by Ben Shneiderman in a recent article he wrote for Issues in Science and Technology [51]. He postulates that human-centered AI will learn from human input and collaboration, assuming AI systems and human will share increasingly more collaboration practices from which they can learn.

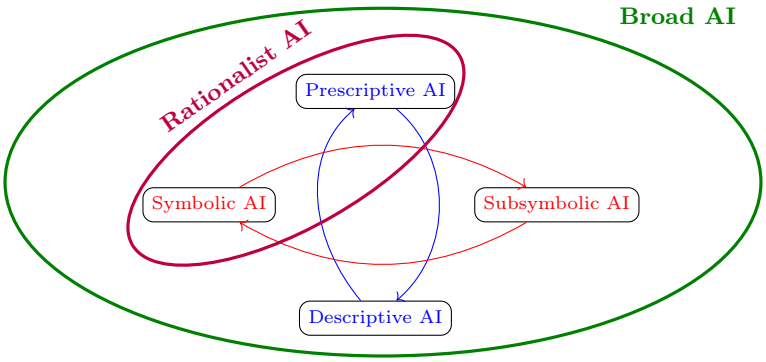
One of the learned lessons, I believe, is that high levels of understanding and control over AI systems are stimulated to the extent that the capabilities of AI systems are aligned with human cognition and abilities, i.e., to the extent that AI systems learn, reason, and act in the world in a similar way as we do. This may be achieved, for example, by ensuring that the features and concepts used to classify, learn, and reason are similar to those used by human. Or, by ensuring that AI systems reason about what they learn. To ensure that the learned knowledge is aligned with their existing knowledge, rational axioms, and norms. Without such an alignment, the decisions of AI systems would be hard, if not impossible, to explain to human users in a comprehensible way.

The human-centered AI community argues against modelling tasks, processes, and procedures beyond human capacity, so that we maintain our ability to treat and evaluate cases individually. Humans should always be able to interrogate and understand why an AI system behaves as it does. This requires that decisions are traceable, that explanations for the made choices are tailored for humans, and that the decisions are aligned with human principles and values, such as consistency, fairness and privacy. In general, the more AI systems behave in accordance with the values and principles that govern human cognition, competences and behaviors, the more natural and effective their collaboration with humans, the lesser the need for adaptation from the human in the loop, and in a sense, the more controllable the AI systems.

The central question is how these values, principles, and standards can be systematically included in the design of AI systems, and how to assess, evaluate

and control their behaviors. In my opinion, a part of the solution is the rationalist view [51] on AI, which is mainly model-driven and reasoning-based using symbolic approaches. Thanks to the rationalist view on AI, we now enjoy a rich literature of formal and computational models of a wide variety of phenomena including reasoning with knowledge, preferences, norms, emotions, planning, decision-making, and natural language processing. These models possess virtues such as explainability and understandability of AI systems. Moreover, the rationalist view on AI comes with rigorous computer science techniques that support traceability. These techniques, for example logical and probabilistic verification methods, causal reasoning, and Bayesian inference, support traceability, which in turn contribute to the controllability of AI systems. In this sense, one may argue that the rationalist view on AI has always been human-centered.

There is now a movement in the AI community that advocates 'broad AI'. It argues in favor of integrating the potentials of the rationalist view on AI with the potentials of modern data-driven methodologies, learning techniques, and sub-symbolic approaches. This would allow us to build powerful AI systems, that are also safe and controllable. In his recently published article with the title "Deep Learning Is Hitting a Wall" [42], Gary Marcus, a cognitive scientist and an AI researcher, advocates this broad view as well. He refers to the recent successes of AI technology such as AlphaGo, where symbolic-tree search and subsymbolic deep learning methods have been combined. Despite many challenges and pitfalls, this broad AI view is intuitive and promising. I believe it has already set an important research direction for the future of AI.



In the rest of this lecture, I will briefly mention some of my research projects, which I believe contribute to this broad AI view.

4 Social and Cognitive Modelling

Some of these projects are the continuation of what we did in the Intelligent Systems group, when it was led by Prof. John-Jules Meyer until 2018. For

example, we are still investigating various logics and formalisms for modelling social and cognitive phenomena such as norms, emotions, and responsibility. Previously, we had developed agent models with emotion related modules (the red part of the diagram in Figure 1; see e.g., [53, 15, 16]). Such an agent can, for example, get in a sad state when it cannot achieve its objective. And, as a consequence of being in a sad state, the agent may drop its unachievable objective and pursue other objectives. This is the recognized role of emotion that make human decision making effective [38, 27, 14, 45]. [1]

We are now extending this line of research by investigating the role of emotions in social settings where agents respond to each other's emotions [41]. This ongoing work is a formalization of the socio-psychological theory proposed by prof. van Kleef on the interpersonal dynamics of emotions [58]. A formalization of this theory can advance human-centered interactions with AI systems. An AI system designed according to such a model would behave differently to an angry human user than to a fearful user.

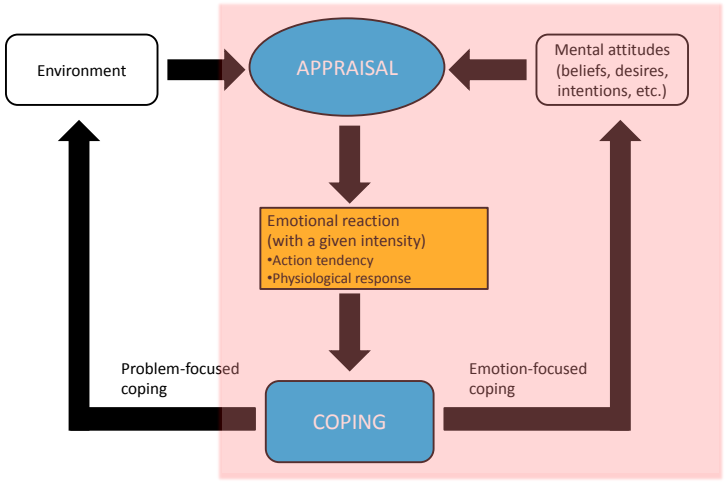


Figure 1: An agent architecture with emotion modules.

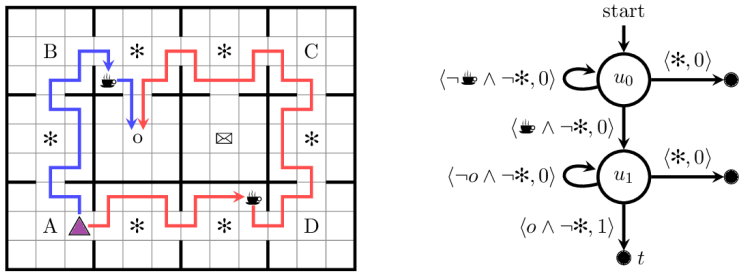
We are also extending our research on norm monitoring and norm enforcement [55, 2, 3, 12, 4, 1, 11, 54, 34] by investigating the data-driven supervision mechanisms that monitor and control the behavior of autonomous systems by collecting their behavioral data to synthesize, revise, and enforce appropriate norms [20, 21, 22, 19]. We use various AI technologies such as logic-based and Bayesian reasoning and optimization techniques. While continuing these

[1] I would like to use this opportunity to thank Prof. John-Jules Meyer for providing me the right platform to do my research in the past two decades; I have learned a lot from working with him, which was always fun and enjoyable.

lines of research, we now investigate foundational and systematic methodologies to integrate model-driven and data-driven approaches, learning and reasoning methods, and symbolic and subsymbolic models. I will explain some of these projects.

5 Logic-guided Reinforcement Learning

A relatively new development in AI, where logic and learning, are combined, is the use of reward-machines in reinforcement learning [40, 56, 33, 32]. In Figure 2, you see a very simple office-like grid environment (taken from [32]). The task of the agent (the triangle) is to deliver coffee to the office (the O-mark). As you can imagine, there are many different ways to perform this task, but only some are optimal (e.g., blue is more optimal than red). Learning to perform this task optimally means that the agent should explore and evaluate lots of possibilities in this environment. This is known as the sample efficiency problem of reinforcement learning. Moreover, if the agent learns to deliver coffee to the office, it cannot use the learned skill to do other related tasks, e.g. the task to go to the office without coffee. This is known as the taskablility problem of reinforcement learning.



(a) The office gridworld

(b) A simple reward machine

Figure 2: An example of a reward machine for an office-like grid environment.

Reward machines can help these two problems by guiding the reinforcement learning agents to learn one subpolicy per subtask. In this example, a reward machine (Figure 2-b) can be used to specify the reward function for the task to deliver coffee to the office by composing two reward functions, one to learn a policy for taking a coffee, and one for learning to go to the office.

Specific challenges in this research area are the automatic synthesis of reward machines for multiagent reinforcement learning, and their decompositions such that individual agents can be trained and can execute the learned policies individually. In a PhD project carried out by Giovanni Varrichione and co-supervised by Prof. Brian Logan and Dr. Natasha Alechina, we investigate automatic synthesis of multiagent reward machines using logical techniques such

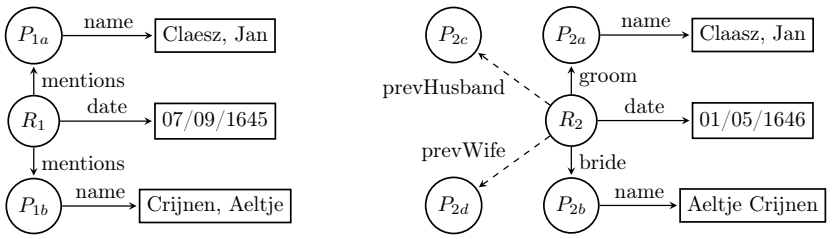
as model checking. The plan is to extend this line of research by automatically synthesising reward machines based on a variety of motivational attitudes including goals, intentions, and norms.

Another recent development in reinforcement learning is the use logical shields to guide the learning process while respecting some safety properties [5, 24, 35]. Shield can help learning agents for example to avoid exploring unsafe actions. We plan to use norms and regulations to devise so-called 'norm-based shields' that govern the exploration activities of agents during training and execution phases. These approaches ensure that the agents learn decision policies that are aligned with rational axioms, or social, legal, and ethical rules.

In two other PhD projects co-supervised by Dr. Shihan Wang, and carried out by Changxi Zhu and Shuai Han, we investigate how communication leads to robust and safe learning in multiagent settings [61]. We study how learning evolves with communication, with the aim of building a theoretical analysis to get a better understanding of the impact of exploiting communication models in multiagent reinforcement learning. These projects investigate various aspects of communication in multiagent reinforcement learning such as the use of specific communication protocols, the use of various communication channels, or the use of communication graphs. In a different research project, carried out by Yangyang Zhao and co-supervised by Dr. Shihan Wang, we focus on dialogue modelling using reinforcement learning. In this project, the sample efficiency problem is tackled using a rule-based teacher model that starts training agents on simple subgoals and gradually train them towards more complex goals.

6 Golden Agents: AI for Digital Humanity

Another project in our group that illustrates the broad AI view is the Golden Agents project. The project aims at searching for patterns that span across decentralized datasets from cultural heritage institutes. To find patterns in such decentralized settings, datasets need to be aligned, which in turn requires identifying duplicate entities within and across various datasets [18]. Recognizing duplicate entities, for example persons, in digital archives is not trivial. As shown in Figure 3, entities may lack attributes (e.g., no birthdate), the values of attributes may be incorrect and incomplete (e.g., names are wrongly spelled), and attribute values are represented using different standards (e.g., using Crijnen, Aeltje and Aeltje Crijnen for the same name).



(a) Burial Record

(b) Prenuptial Marriage Record

Figure 3: Each registry in the Amsterdam City Archives has records $(R_1, \ldots, R_n)$, all of which are associated with one or more references to persons $(P_{i1}, \ldots, P_{ik})$. In this figure we show a simplified view, where URI nodes are drawn as circles and literal nodes as rectangles. In this example, we might deduce that nodes P_{2a} and P_{1a} represent the same person. The same goes for nodes P_{2b} and P_{1b}. (taken from [6])

To find duplicate entities, Jurian Baas, Dr. Ad Feelders and I, investigate possible integrations of subsymbolic models, generated by embedding techniques, with symbolic techniques and domain knowledge [7, 6]. We work with knowledge graphs containing entities that represent persons. Entities are first encoded using their context information from the graph and then embedded as vectors in a multidimensional Euclidean space. The proximity of vectors is then used to measure the similarity between the corresponding entities. The more similar the context of two entities, the more similar their encoding vectors, and thus the more proximate the vectors in Euclidean space. We then apply clustering techniques to find pairs of entities that are likely duplicates, but this should be done with care. Since we are searching for duplicates, we need to ensure that the duplicates found, satisfy structural properties of the same-as relation such as transitivity. If entities x and y are found as duplicates, and y and z as well, then you may expect that x and z are found as duplicate too. We have shown that ad-hoc application of transitivity rule to the elements of the found clusters may incorrectly identify duplicates and thereby decrease precision. We have used symbolic techniques to prevent such false positives. This process of finding duplicate entities is shown in Figure 4. The combination of sub-symbolic embedding with symbolic graph editing techniques and domain knowledge has shown to improve the performance compared to ad-hoc application of the embedding technique [8]. We are currently using similar AI technologies to detect communities of individuals that were involved in the creative industry in the Netherlands during the Dutch Golden Age.

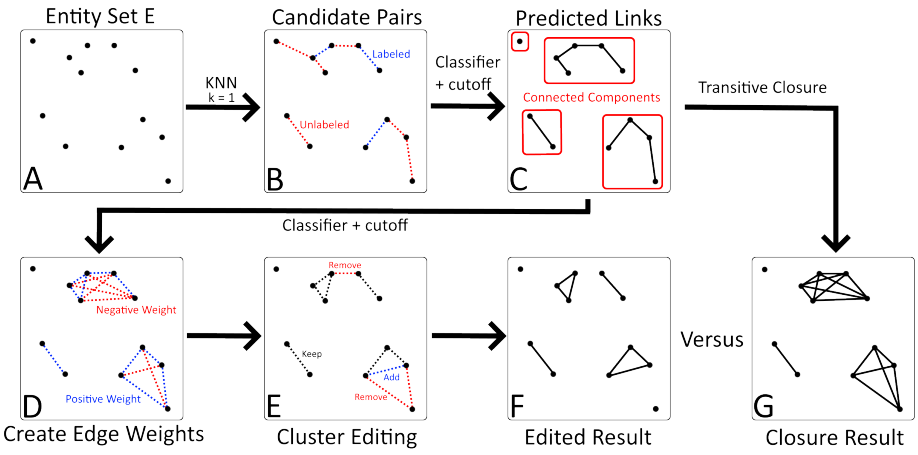


Figure 4: The process of finding duplicate entities.

7 Reasoning with Uncertainty for Responsibility and Causality

As I mentioned earlier in this lecture, formal and computational methods that allow us to trace the underlying reasoning and the route of the decisions made by the AI systems, are essential for their controllability. Such mechanisms can be used, for example, to trace the cause of accidents where AI systems are involved and to determine who are responsible for the caused accidents. Think of traffic accidents where self-driving cars are involved, or the so-called 'flash crash' incidents in financial markets where algorithmic traders are involved [52].

Previously, we have proposed formal models for reasoning about responsibility [10, 59, 60]. According to these models, an intelligent agent (AI system or human) is held responsible for an accident if it was able to prevent the accident and was aware of it [26]. To explain the notion of responsibility and our approach for modelling it consider the example, known as the Traveler and Two Enemies, proposed by James Angell McLaughlin [43]. Imagine a traveler P who needs water to survive a trip across the desert. The traveller P has two enemies $E1$ and $E2$. The night before traveller's departure, and while the traveller was sleeping, $E1$ adds poison to the water in the traveller's canteen. Later, while the traveller still sleeps, the $E2$ empties the (poisoned) water from the canteen. The traveller dies of thirst in the middle of the desert. The question is which of the two enemies and to what extent are responsible for the death of the traveller. This situation can be formally modelled by the graph presented in Figure 5. The path enclosed by the blue line represent the history I just explained. i.e., the first enemy adds poison to the canteen and the second enemy empties the

canteen. The other paths indicate alternative possibilities. The history above informs us that none of the enemies tried their alternative possibilities to save the traveller. Our proposed analysis identifies that both enemies are in this case responsible with equal degree (for details see [59]).

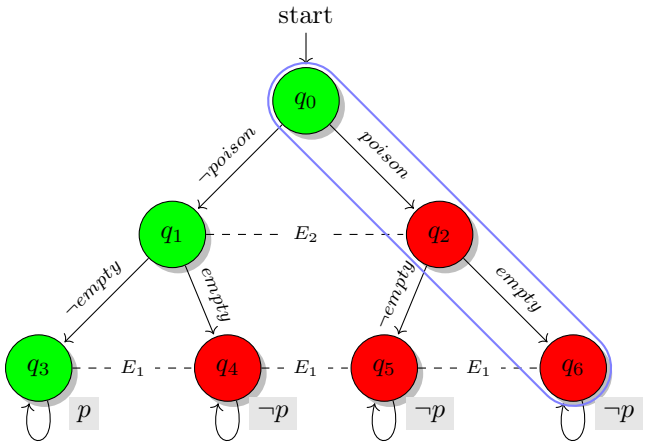


Figure 5: In q_0, E_1 may poison the water or not. In q_1 and q_2, E_2 can either empty the canteen or not. As a result, P is alive in q_3 (represented by proposition p) and dead in q_4, q_5, and q_6 (represented by $\neg p$). The path outlined in blue denotes the history.

In recent AI literature, Chockler and Halpern [13, 29] takes a similar view and propose the notion of a degree of responsibility and blame, where they use the Judea Pearl's definition of causality for the case of a single agent. We are currently working on modelling responsibility in probabilistic contexts where agents can be held responsible, not necessarily for realising a bad outcome, but also for increasing the probability of the bad outcome. Not following traffic rules does not necessarily cause an accident, but it may increase the probability of an accident to an unacceptable level. We have several ongoing projects that contribute to this traceability theme, and I would like to mention some of them.

In two projects carried out by Maksim Gladyshev and Nima Motamed, and co-supervised by Dr. Natasha Alechina and Dr. Dragan Doder, we investigate how to reason about probability changes. Moreover, in the CAUSES project funded by NWO and ProRail, we investigate learning and reasoning with causal models to identify and trace the causal relationship between the local decisions of AI systems and the overall system behavior. The CAUSES project is carried out by Kristina Gogoladse and Francisco Simoes , co-supervised by Dr. Natasha Alechina, Prof. Brian Logan, Dr. Thijs van Ommen, myself and of course our ProRail colleagues Emdzad Sehic and Wilco Tielam. Finally, as a part of the gravitation project 'Hybrid AI', Annet Onnes, co-supervised by

Dr. Silja Renooij and myself, investigates monitoring mechanisms for tracing deviant and uncertain behaviors of AI systems.

8 Large-scale Data-Driven Agent-based Simulations

The final project that I would like to mention integrates data- and model-driven approaches in building an agent-based simulation framework. Agent-based simulation is an AI-technology that can be used to conduct computational experiments to analyze and predict complex social phenomena such as the evolution of economic inequality, seasonal migrations, traffic, and the spread of diseases. However, state-of-the-art agent-based simulation frameworks do not scale to the behavioural and interaction complexity, and the large numbers of agents required for realistic computer simulations of social systems [37, 36]. To overcome these limitations, Jan de Mooij, Dr. Davide Dell'Anna, Prof. Brian Logan, Dr. Samarth Swarup and Dr. Parantapa Bhattacharya and myself, have developed a scalable agent-based simulation framework to support the computational modelling of complex social systems. A key characteristic of this newly developed simulation framework is its data-driven design [9].

This simulation framework is used to study the effects of non-pharmaceutical intervention such as mask wearing, social distancing, and school closures, in the fight against the spread of COVID-19 [17]. We used available data sources from various counties of the US state of Virginia. The experiments are scaled to the entire population of Virginia ($\sim$8 million agents). The overall setting of these simulation experiments is illustrated in Figure 6. The input to the simulation consists of a synthetic population with realistic demographics, weekly activity schedules, political orientations, and activity locations drawn from real locations and building data sets. In the chosen counties, the number of individuals ranges from 20k to 180k and the number of weekly visits to locations ranges from 680k to about 6 million. Each individual in the synthetic population is represented by a software agent that reasons about its beliefs including its sense information (e.g., the number of symptomatic individuals the agent has observed previously), its objectives (e.g., daily activities), and its trust in government determined by its political orientation, to decide whether to comply with the non-pharmaceutical interventions such as maskwearing and social distancing that were introduced in Virginia between March and July 2020.

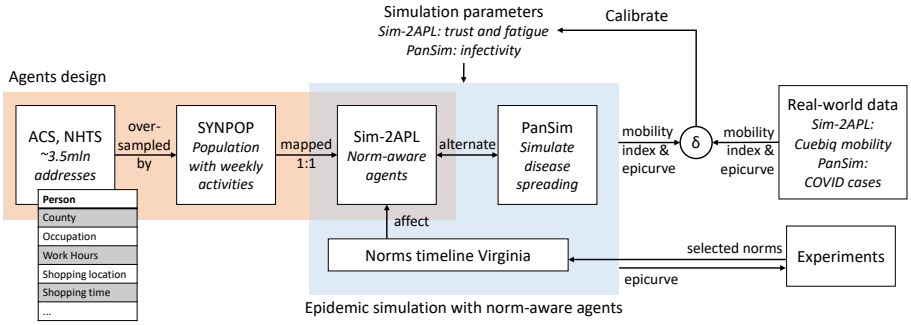


Figure 6: The overall setting of the agent-based COVID-19 simulation.

In this simulation-experiment, each individual in the synthetic population is modelled by a software agent that decides to comply with, or to violate, the non-pharmaceutical interventions. The decision rule of each agent uses a variety of parameters such as the number of symptomatic individuals that the agent has observed previously, its scheduled daily activities, and its trust in the government. The activities decided by the agents are then sent to a disease model which determines the progression and spread of the virus. The model-driven rule-based decision models of the agents are then calibrated to the anonymized cellphone-based mobility data. The result is illustrated in Figure 7. As shown, we could closely simulate the general mobility pattern in a certain period of time where non-pharmaceutical interventions have been in place. Currently, we use the calibrated model to investigate various COVID-related non-pharmaceutical interventions to find out what could be learned from the type and timing of these interventions.

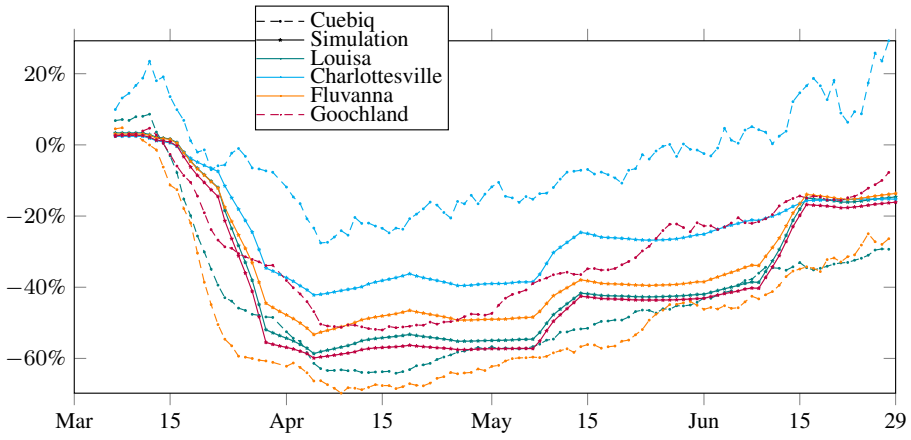


Figure 7: The calibration result of the COVID-19 simulation.

9 Concluding remarks

It is now time to conclude my lecture with some final remarks. In the relatively short history of Artificial Intelligence, we have been witnessing several AI-summers and AI-winters. These periods were characterized by the emergence or the absence of AI progress and innovations in both science and technology, which in turn, have led to an increase or decrease of AI funding and investments [25]. We are now in a new AI-summer period where hype, enthusiasm and optimism surrounding AI has peaked again and expectations are, I would say, quite high. These are all good news, but there is also a down-side to the hype around AI that may push us into a new AI-winter again. The optimism and enthusiasm in a scientific field is a great thing, but the AI community and in particular the academic AI needs to be cautious not to create unrealistic expectations and promises.

Also, the current polarization within the AI community is in my opinion not a constructive development. Taking an ideological stance on sub-symbolic AI, as currently advocated by Geoffrey Hinton (the grand, grand, grandson of the famous logician Goerge Boole), forces us to ignore existing scientific knowledge concerning intelligent phenomena (by the way, all obtained by small data) and to re-learn them altogether from scratch (probably with big data) (see also [42]). Existing scientific knowledge, which is available in symbolic form, has proven to be helpful for developing understandable, explainable and traceable AI models, and I do not see any reason not to use them. Similarly, an ideological stance on symbolic AI would ignore powerful and effective techniques that allow us to model intelligent phenomena such as sensory perception, or to work with imprecise and faulty data. Thanks to these techniques, many tasks and innovations, which were previously not even thinkable, are now possible.

The few research projects that I just presented are some examples of how these extremes may complement and strengthen each other. Together with the rest of the Intelligent System group, we will continue exploring systematic and effective integrations of AI techniques. We will do our best to contribute to the broad AI vision that can support the design and development of powerful yet controllable AI systems.

10 Besluit

Aan het eind van mijn rede gekomen, wil ik graag het college van bestuur van de Universiteit Utrecht, de Faculteit Bètawetenschappen, het departement Informatica, en allen die mijn benoeming hebben gesteund bedanken voor het in mij gestelde vertrouwen. Mijn speciale dank gaat uit naar mijn collega's met wie ik met veel plezierig en productief heb samengewerkt. Ik heb al een aantal namen tijdens mijn presentatie genoemd, maar er zijn er veel meer die ik, gezien de tijd, niet een voor een kan noemen. Een aantal namen staan op de slide en

ik wil ze nogmaals bedanken voor de samenwerking en vriendschap.

Ook wil ik mijn dank betuigen aan de collega's met wie ik de afgelopen jaren samengewerkt heb aan het doorbreken van barrières tussen wetenschappelijke disciplines en faculteiten, om een sterk interdisciplinaire AI-onderzoek en AI-opleidingen in Utrecht neer te zetten. Samen met Pinar Yolum, Floris Bex, Rosalie Iemhoff, Chris Jansen, Tejaswini Deoskar en Martijn Mulder hebben we de afgelopen 5 jaar hard gewerkt aan het verbeteren van de AI-master opleiding. De resultaten mogen er zijn. De Nationale studenten enquêtes van de afgelopen 3 jaren laten zien dat studententevredenheid gestaag gegroeid is. Daar zijn we allemaal erg trost op.

Op het onderzoeksgebied hebben ik samen met Jan Broersen, Yoad Winter, Stefan van der Stigchel, Henk Aarts, en de coördinatoren Sara Simmerlink en Hanneke Roodbeen, gewerkt aan het bouwen van een interdisciplinaire AI-gemeenschap aan Universiteit Utrecht. Dit hebben we gedaan door het inrichten van het Human-centered AI-focusgebied welke inmiddels 10 Special Interest Groups rijk is. Mijn dank gaat uit naar de trekkers van deze special interest groups; ze hebben fantastisch werk gedaan om de AI-community op universiteit Utrecht bij elkaar te brengen.

Om de fundamentele relatie met digitalisering en data science te bevorderen werken we nu ook nauw samen met twee andere aangrenzende focusgebieden Governing the Digital Society onder leiding van Prof. Jose van Dijck, Prof. Anna Gerbrandy, Prof. Nadya Purtova, Prof. Janneke Gerard, prof. Albert Meyer, en dr. Mirko Schaefer, en Applied Data Science onder leiding van Prof. Peter van der Heijden. Het is een groot plezier te mogen werken met deze collega's. Ook wil ik graag mijn collega's in het departement informatica en faculteit Exacte Wetenschappen bedanken voor hun inzet en enthousiasme voor Kunstmatige intelligentie.

De in Utrecht gevestigde organisatie UAF wil ik hartelijk bedanken. Ze ondersteunen vluchtelingen om zich te kunnen ontwikkelen op het gebied van studie en werk. Zonder hun steun in 1985, toen ik als vluchteling naar Nederland kwam, had ik waarschijnlijk mijn wetenschappelijke carrière niet kunnen maken.

Tot slot wil ik mijn vrouw, Cathalijne Smulders, mijn super lieve dochter Aya, en de rest van mijn familie in Iran en Nederland bedanken voor hun steun. Cathalijne, je hebt me afgelopen jaren enorm geholpen. We hebben lange, heel interessante discussies gehad over wetenschap en kunst, over vragen die in mijn werk tegenkom, ook de wetenschappelijke. Iedere keer als ik dacht dat ik het eindelijk goed had, kwam je met een reflectie dat het toch anders kan zijn. En Aya, je zet me altijd aan het denken met je herhaaldelijke vraag "Papa, hoe zo zegt je dit". Ik hoop dat je altijd vragen blijft stellen.

Ik heb gezegd!

References

[1] Natasha Alechina, Nils Bulling, Mehdi Dastani, and Brian Logan. Practical run-time norm enforcement with bounded lookahead. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015*, pages 443–451. ACM, 2015.

[2] Natasha Alechina, Mehdi Dastani, and Brian Logan. Programming norm-aware agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, pages 1057–1064, 2012.

[3] Natasha Alechina, Mehdi Dastani, and Brian Logan. Reasoning about normative update. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI 2013*, pages 20–26. IJCAI/AAAI, 2013.

[4] Natasha Alechina, Mehdi Dastani, and Brian Logan. Norm approximation for imperfect monitors. In *Proceedings of the 13th International conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2014*, pages 117–124. IFAAMAS/ACM, 2014.

[5] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2669–2678. AAAI Press, 2018.

[6] Jurian Baas, Mehdi Dastani, and Ad Feelders. Tailored graph embeddings for entity alignment on historical data. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, pages 125–133, 2020.

[7] Jurian Baas, Mehdi M Dastani, and Ad J Feelders. Entity matching in digital humanities knowledge graphs. In *Proceedings of the Second Conference on Computational Humanities Research (CHR2021)*, pages 1–15, 2021.

[8] Jurian Baas, Mehdi M Dastani, and Ad J Feelders. Exploiting transitivity for entity matching. In *Proceedings of the European Semantic Web Conference*, pages 109–114. Springer, 2021.

[9] Parantapa Bhattacharya, Jan de Mooij, Davide Dell'Anna, Mehdi Dastani, Brian Logan, and Samarth Swarup. PanSim + Sim-2APL: A platform for large-scale distributed simulation with complex agents. In *Proceedings of the 9th International Workshop on Engineering Multi-Agent Systems, EMAS@AAMAS 2021*, volume 13190, pages 1–21, 2021.

[10] Nils Bulling and Mehdi Dastani. Coalitional responsibility in strategic settings. In João Leite, Tran Cao Son, Paolo Torroni, Leon van der Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems - 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings*, volume 8143 of *Lecture Notes in Computer Science*, pages 172–189. Springer, 2013.

[11] Nils Bulling and Mehdi Dastani. Norm-based mechanism design. *Artificial Intelligence*, 239:97–142, 2016.

[12] Nils Bulling, Mehdi Dastani, and Max Knobbout. Monitoring norm violations in multi-agent systems. In *Proceedings of the 12th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2013*.

[13] Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[14] Antonio R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam, 1994.

[15] Mehdi Dastani and Emiliano Lorini. A logic of emotions: from appraisal to coping. In *AAMAS*, pages 1133–1140, 2012.

[16] Mehdi Dastani, Emiliano Lorini, John-Jules Meyer, and Alexander Pankov. Other-condemning anger= blaming accountable agents for unattainable desires. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 15–33. Springer, 2017.

[17] Jan de Mooij, Davide Dell'Anna, Parantapa Bhattacharya, Mehdi Dastani, Brian Logan, and Samarth Swarup. Quantifying the effects of norms on COVID-19 cases using an agent-based simulation. In *Proceedings of the 22nd International Workshop on Multi-Agent Systems and Agent-Based Simulation, MABS@AAMAS 2021*, volume 13128, pages 99–112, 2021.

[18] Jan de Mooij, Can Kurtan, Jurian Baas, and Mehdi Dastani. A computational framework for organizing and querying cultural heritage archives. *Journal of Computing and Cultural Heritage*, sep 2021.

[19] Davide Dell'Anna, Natasha Alechina, Brian Logan, Maarten LÃ¶ffler, Fabiano Dalpiaz, and Mehdi Dastani. The complexity of norm synthesis and revision. In *Proceedings of the 15th International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems, COINE@AAMAS 2022 (To Appear)*, 2022.

[20] Davide Dell'Anna, Fabiano Dalpiaz, and Mehdi Dastani. Validating goal models via bayesian networks. In *Proceedings of the 5th International Workshop on Artificial Intelligence for Requirements Engineering, AIRE@RE 2018*, pages 39–46, 2018.

[21] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime Norm Revision using Bayesian Networks. In *Proceedings of the 21st International Conference on Principles and Practice of Multi-Agent Systems (PRIMA2018)*, 2018.

[22] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of norms and sanctions based on agent preferences. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2019*, pages 1609–1617, 2019.

[23] Hubert L. Dreyfus. *What Computers Still Can't Do.* MIT Press, revised edition edition, 1992.

[24] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. AAMAS 2021, pages 483–491, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

[25] E. Francesconi. The winter, the summer and the summer dream of artificial intelligence in law. *Artificial Intelligence and Law*, 30:147–161, 2022.

[26] Harry G Frankfurt. Alternate possibilities and moral responsibility. *The journal of philosophy*, 66(23):829–839, 1969.

[27] Nico H Frijda. Emotions and action. In *Feelings and emotions: The Amsterdam symposium*, pages 158–173, 2004.

[28] J. Gerards, M.T. Schäfer, Vankan A., and I. Muis. Impact assessment mensenrechten en algoritmes. `https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes`, 2021.

[29] Joseph Y. Halpern. *Actual Causality.* The MIT Press, 2016.

[30] High-Level Expert Group on Artificial Intelligence. A definition of AI: Main capabilities and scientific disciplines. `https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf`, 2019.

[31] Douglas R Hofstadter. *Gödel, Escher, Bach: an eternal golden braid.* Penguin books. Basic Books, New York, NY, 1979.

[32] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

[33] Kishor Jothimurugan, Suguman Bansal, Osbert Bastani, and Rajeev Alur. Compositional reinforcement learning from logical specifications, 2021.

[34] Max Knobbout and Mehdi Dastani. Reasoning under compliance assumptions in normative multiagent systems. In *Proceedings of the 11th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2012*, pages 331–340, 2012.

[35] Bettina Könighofer, Florian Lorber, Nils Jansen, and Roderick Bloem. Shield synthesis for reinforcement learning. In Tiziana Margaria and Bernhard Steffen, editors, *Leveraging Applications of Formal Methods, Verification and Validation: Verification Principles*, pages 290–306, Cham, 2020. Springer International Publishing.

[36] Averill M. Law. *Simulation Modeling and Analysis*. New York: Mcgraw-Hill, 2015.

[37] D.R. Law. Scalable means more than more: a unifying definition of simulation scalability. In *1998 Winter Simulation Conference. Proceedings (Cat. No.98CH36274)*, volume 1, pages 781–788 vol.1, 1998.

[38] Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

[39] Hector J. Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[40] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. of the 11th International Conference on International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann Publishers Inc., 1994.

[41] Jieting Luo and Mehdi Dastani. Modeling affective reaction in multi-agent systems. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor, editors, *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, pages 1681–1683. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.

[42] Gary Marcus. Deep learning is hitting a wall. `https://nautil.us/deep-learning-is-hitting-a-wall-14467/`, 2022.

[43] James Angell McLaughlin. Proximate cause. *Harvard law review*, 39(2):149–199, 1925.

[44] Peter Norvig. On chomsky and the two cultures of statistical learning. `http://norvig.com/chomsky.html`, 2011.

[45] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[46] Corien Prins, Haroon Sheikh, Erik Schrijvers, Eline de Jong, Monique Steijns, and Mark Bovens. Mission AI: The New System Technology. `https://english.wrr.nl/publications/reports/2021/11/11/summary-mission-ai`, 2021.

[47] Howard Raiffa, Amos Tversky, and David E. Bell. *Decision making : descriptive, normative, and prescriptive interactions*. Cambridge University Press, 1988.

[48] Stuart J. Russell and Peter Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.)*. Pearson Education, 2010.

[49] John Searle. Why dualism (and materialism) fail to account for consciousness. In Richard E. Lee, editor, *Questioning Nineteenth Century Assumptions About Knowledge, Iii: Dualism*, pages 5–48. Suny Press, 2010.

[50] John R Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.

[51] B. Shneiderman. Human-centered ai. *NAS Issues in Science and Technology*, 37(2):56–61, Winter 2021.

[52] Ian Sommerville, Dave Cliff, Radu Calinescu, Justin Keen, Tim Kelly, Marta Kwiatkowska, John Mcdermid, and Richard Paige. Large-scale complex it systems. *Communication ACM*, 55(7):71–77, 2012.

[53] Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. A formal model of emotion triggers: an approach for BDI agents. *Synth.*, 185(Supplement-1):83–129, 2012.

[54] Bas Testerink, Mehdi Dastani, and Nils Bulling. Distributed controllers for norm enforcement. In *Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI 2016*, volume 285, pages 751–759. IOS Press, 2016.

[55] Nick A. M. Tinnemeier, Mehdi Dastani, John-Jules Ch. Meyer, and Leendert W. N. van der Torre. Programming normative artifacts with declarative obligations and prohibitions. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009*, pages 145–152, 2009.

[56] Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. Learning reward machines for partially observable reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 15497–15508. Curran Associates, Inc., 2019.

[57] A. M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 10 1950.

[58] Gerben A Van Kleef. *The interpersonal dynamics of emotion*. Cambridge University Press, 2016.

[59] Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. Strategic responsibility under imperfect information. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 592–600. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[60] Vahid Yazdanpanah, Enrico H. Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M. Jonker, and Timothy J. Norman. Responsibility research for trustworthy autonomous systems. In Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé, editors, *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, pages 57–62. ACM, 2021.

[61] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent reinforcement learning with communication. *CoRR*, abs/2203.08975, 2022.