Abstract

Search-based AI agents are state of the art in many challenging sequential decision-making domains. However, contemporary approaches lack the ability to explain, summarize, or visualize their plans and decisions, and how they are derived from traversing complex spaces of possible futures, contingencies, and eventualities, spanned by the available actions of the agent. This limits human trust in high-stakes scenarios, as well as effective human-AI collaboration. In this paper, we propose and motivate the new research direction of explainable search. We discuss its differences to existing approaches in explainable AI, and outline important related research challenges with concrete examples, focusing in particular on online interactions and the resulting understanding of explanations in an ongoing process of mutual collaboration towards human goals.

1 Introduction

In recent years, AI has increasingly found its way from research labs into applications: from the recommendation systems used by online retailers to image recognition on social networks, from voice-controlled personal digital assistants to medical diagnosis systems, from service robots to self-driving vehicles. As we work with AI and rely on AI for more and more decision-making processes that influence our daily lives, issues around user understanding of such processes have garnered attention. Aimed at goals such as supporting trust in AI, increasing user satisfaction with AI, enhancing collaboration with AI, and enabling transparency of AI decision-making—and partly also motivated by new European Union regulations on a “right to explanation” [Goodman and Flaxman, 2017]—the research area of explainable AI (XAI) has rapidly developed.

When the DARPA XAI program [Gunning, 2017] helped kickstart explainable AI, the core focus of the field was on explainable machine learning, as the “black-box” properties of the suddenly ubiquitous deep neural networks were seen as a central problem for understandability or interpretability of AI systems. This is probably still the most well-developed subfield of XAI, with a variety of surveys covering recent progress [Samek et al., 2017; Adadi and Berrada, 2018; Guidotti et al., 2019; Henin and M étayer, 2019]. While there has been a strong focus on explaining single algorithmic decisions of data-driven systems such as neural networks, the challenge of explaining complex behavior of goal-driven systems, i.e. agents autonomously acting in their environment through sequences of decisions, has only recently come more into view [Anjomshoae et al., 2019; Sado et al., 2020]. Multiple authors have observed that much of the existing XAI literature is not suitable for explaining sequences of decisions [Topin and Veloso, 2019] and “studies addressing the increasingly pervasive goal-driven agents and robots are still missing” [Lage et al., 2019], even though “sequential environments offer a unique challenge for generating human understandable explanations” [Ehsan et al., 2019]. In such contexts, the decisions to be explained could for example be part of both short- and long-term plans to achieve goals at different levels of granularity, and could require explanation both in advance of execution to manage expectations, during execution to explain deviations or unforeseen problems, and after execution for debriefing purposes [Gervasio et al., 2018].

Looking at the existing work on explaining goal-driven behavior, a large part of it falls into the emerging subfield of explainable planning [Fox et al., 2017; Chakraborti et al., 2019a; Chakraborti et al., 2020], which focuses on the classical planning setting. For reinforcement learning (RL) however, even though it is arguably at the heart of many recent AI breakthroughs [Mnih et al., 2015; Silver et al., 2016; Silver et al., 2017c; Silver et al., 2017b; Yang et al., 2019; Berner et al., 2019; Vinyals et al., 2019], work towards explainability is just beginning, and remains one of the major research challenges [Dulac-Arnold et al., 2019]. The existing work on explainable reinforcement learning (XRL) often either attempts to explain an agent’s behavior in its entirety, i.e. tries to simplify, characterize, or summarize a complete and fully optimized agent policy [Hayes and Shah, 2017; Verma et al., 2018; Amir and Amir, 2018; Roth et al., 2019];

We understand the term RL to also cover planning, as learning from simulated experience with the help of a model [Sutton and Barto, 2018]. While this blurs the lines between RL and classical planning problems, we consider the RL setting to be more relevant for the purposes of this paper, as it typically focuses less on solvability of problems or optimality of plans found offline, and lends itself more naturally to online decision-making, in particular in potentially stochastic, partially observable, and/or multi-agent domains.
or it focuses on explaining a single agent decision instead, i.e. tries to illustrate how and why a given policy maps a specific state to a specific action [Khan et al., 2009; Iyer et al., 2018; Ehsan et al., 2018; Erwig et al., 2018]. In many application scenarios however, a high-level explanation of an entire policy is likely to be either too abstract or too verbose to be helpful – imagine a chess tutor agent trying to explain to a human student how it plays chess in general, instead of explaining its next move recommendation to increase pressure on the opponent’s rook; or a search & rescue robot explaining to a human mission supervisor the general structure of its decision making process, instead of outlining its plan for getting access to the first floor of the partially collapsed building it is in. A myopic explanation of a single decision, without explicitly discussing the future situations and further decisions it might lead to, would not give the user enough relevant information for these settings either – imagine the chess tutor simply highlighting that the upper left corner of the board contained the most relevant input features for its move decision, or the search & rescue robot only stating that moving left leads to an 11% higher probability of clearing the first floor within the next ten minutes compared to moving right. Humans would probably want to know: Why? What exactly could happen next? Which possible outcomes were explored, how were they interpreted, compared, and selected from?

Learning complete high-quality policies for the entire state space in complex domains can be very challenging. In cases where domain models are either available or can be learned, the most promising RL approach is often online planning, or the repeated search for the best next action, starting from the current state in which a decision is needed, and exploring possible futures in the immediate future of the agent [Silver et al., 2016; Silver et al., 2017c; Silver et al., 2017b; Schrittwieser et al., 2019; Segler et al., 2017]. The outcomes of these searches are neither complete plans to solve the entire task, nor single actions in isolation. They are complex trees of expected contingencies and eventualities, starting from the here and now, together with the agent’s current best idea of how to handle them, and finally motivating a most promising next action to execute before replanning. In this paper, we call for research on how to explain the decision-making of such search algorithms, which is best framed neither as full-policy nor as single-action explanation, but as exploration of possible futures, their evaluations, their relationships to each other, and the available choices between them.

With explainable search, we therefore propose a direction of XAI that is putting a different, so far underappreciated “black box” into its center – not the large graphs of elements such as neurons, layers, weights, and activation functions that make up neural networks, but the large graphs of elements such as actions, states, observations, and rewards that make up search trees. While tree structures such as decision trees have traditionally been seen as interpretable, this strongly depends on their complexity [Arrieta et al., 2020], and a typical search tree “of possible futures is a large object with many potential branches that is difficult to understand even for sophisticated users” [Dodson et al., 2011]. Debugging tree or graph structures with thousands of nodes and their connections, often annotated with various statistics collected and estimated, aggregated and processed during search, is challenging even for the search algorithm designer, and understanding them in real time is impossible for end users such as the chess student or the search & rescue supervisor mentioned before. For this reason, we believe that while the challenge of explainable AI has long been understood to cover explaining both “why AlphaGo selected a specific move at each turn, or on what basis a neural network recognises an image as an ‘image of a cat’” [Fox et al., 2017], the former problem has long been neglected in favor of the latter.

This paper poses and explores the research challenge of explainable search. Section 2 discusses related work; Section 3 outlines some of the particular challenges and properties of explainable search; and Section 4 summarizes and concludes.

2 Related Work

We envision research that is inspired by, but clearly distinct from, the following two main strands of work on explainable reinforcement learning.

Explainable search is distinct from existing work on explaining entire RL policies. This includes work aiming at learning an interpretable secondary policy to approximate an uninterpretable primary policy [Verma et al., 2018; Hein et al., 2018; Topin and Veloso, 2019; Lee, 2019; Koul et al., 2019]; work aiming at learning an interpretable primary policy from scratch [Roth et al., 2019]; work aiming at explaining the inner workings of an uninterpretable policy [Zahavy et al., 2016; Sreedharan et al., 2020b]; and work aiming at giving an overall impression of a policy by providing typical behavior examples, or identifying key moments of its interactions [Huang et al., 2018; Amir and Amir, 2018; Amir et al., 2018; Lage et al., 2019].

Explainable search does not take an entire policy into view, but instead mainly an agent’s behavior in its current situation and the currently possible, expected, or desired situations in its near future, as represented in its search tree. We expect this focus to result in more relevant content for many types of explanations; and we expect these focused explanations to be able to provide more detail, while still remaining more cognitively manageable for the human user – compare explaining the preferred next chess move to explaining how to play chess.

Explainable search is also distinct from existing work on explaining individual decisions of non-searching RL agents. By this we mean research that aims at providing information on how or why a given decision was made, but without referring explicitly to different possible futures and subsequent decisions this decision can lead to, or its alternatives could have led to, for the agent. This includes techniques that work by highlighting the most relevant variable for the decision in a factored MDP context [Elizalde et al., 2007; Elizalde et al., 2008]; by identifying which high-reward or low-reward states a decision could ultimately lead to, but without detailing how [Khan et al., 2008; Khan et al., 2009]; by learning a translation from agent decisions to a corpus of human behavior explanations for the same domain [Ehsan et al., 2018]; by decomposing expected rewards into multiple components that carry their own semantics [Erwig et al., 2018; Pocius et al., 2019]; or by using saliency maps for the visual-
ization of deep neural network behavior [Iyer et al., 2018].

While these techniques can provide interesting insights in summary form and at a relatively high level of abstraction, due to not conducting any search none of them are able to explicitly refer to the space of possible futures spanned by the agent’s legal actions and their consequences. Explaining this space to the user, and exploring it together with the user, is something explainable search is uniquely tailored for.

Both previous work on explaining entire policies, and previous work on explaining individual decisions in isolation – effectively taking snapshots of entire policies – suffers from the fact that non-searching “RL agents do not need to plan or reason about their future to select actions, which makes it hard for them to explain their behavior — all they know is that they should perform a particular action in a situation, in the case of deterministic policies, or select an action according to a probability distribution, in the case of stochastic policies. The ‘why’ behind decision-making is lost during the learning process as the policy converges to an optimal action-selection mechanism. At most, agents know that choosing one action is preferable over others, or that some actions are associated with a higher value — but not why that is so or how it came to be.” [Sequeira et al., 2019] Some explainable RL approaches try to recover small parts of this “why behind decision-making” with the help of extra bookkeeping during learning, or during trajectories specifically simulated in order to derive policy explanations [van der Waa et al., 2018; Sequeira et al., 2019; Cruz et al., 2019]. Compared to search-based explanations however, their resulting explanations are not very rich and flexible in content, and e.g. only represent summary probabilities or frequencies of the specific events tracked by the proposed bookkeeping, as expected under the learned policy, or an alternative policy proposed by the user. In contrast, explainable search can retain and use the entire search tree of the underlying search algorithm for explanation purposes, and thus explicitly explain decisions by reasoning about different possible futures, the probabilities of different future events, and the behavior currently estimated to be most promising by the search in different possible scenarios deemed relevant for the user’s understanding.

Note that while the scope of explainable search lies in a sense between that of previous work on explaining individual decisions in isolation, and previous work on explaining entire policies, we argue that the research challenge of explainable search goes beyond both of these subfields. Most previous work on explainable RL, for example assumes that the decision or policy to be explained has already been fully optimized by a learning algorithm; the behavior is assumed to be optimal, or at least final, before the explanation process begins. Contrarily, the plans and the domain understanding of an agent using search are always under construction—they are not only changing and evolving during the search for the next action, but also from search to search, from timestep to timestep throughout any given episode, such as an ongoing game of chess or an ongoing rescue mission (and potentially from episode to episode as well, if learning is involved). This means that explainable search has to be able to handle the online, sequential nature of the decision-making, ongoing tasks as well as ongoing user interactions and ongoing needs for communication and explanations; changes in plans, surprising events or obstacles, and its own (or the human’s) potential mistakes and corresponding revelations or revisions will have to be processed by an explainable search agent, as discussed in more detail in Section 3 below.

Explainable search has connections to or overlap with a number of other research areas. These include for example visualizations of heuristic search [Magnaguagno et al., 2017], which typically do not consider the online setting we are focusing on here, and only provide limited information such as the overall shape of the search tree that was needed to fully solve a problem, and the heuristic evaluations of states within. The only work on explanations in an online RL setting to the best of our knowledge, and the closest related work to ours, is using a “bounded lookahead procedure” in every timestep, instead of solving for a complete policy [Wang et al., 2016]. However, this lookahead procedure only covers the immediate next decision, not multiple timesteps into the future, and is applied to a test domain with strongly limited numbers of actions, beliefs, and possible outcomes. Its explanations are therefore limited when compared to those envisioned here for large search trees in complex domains, but it could be thought of as a first stepping stone into the direction we propose.

Many recent contributions to explainable AI are general enough to be relevant to explainable search as well: for example the notion of an interpretability-completeness tradeoff for explanations [Gilpin et al., 2018], work on different modes of interpretability evaluation [Doshi-Velez and Kim, 2017], and insights from the social sciences on what constitutes a good explanation [Miller, 2019], to name only a few. We believe that there is great value in understanding explainable search as an interdisciplinary effort as well, in order to integrate different perspectives on related concepts such as explanation, advice, argumentation, storytelling, visualization and verbalization, multi-agent systems, and human-AI collaborative systems in general.

3 Challenges of Explainable Search

In this section, we highlight six research challenges that are of particular importance to explainable search. We relate them to and motivate them with the examples introduced in Section 1: the chess tutor agent interacting with its student, and the search & rescue robot interacting with its supervisor.

Explanations as conversations. Several authors have found that a successful explanation can require more than a single transfer of information from the explainer to the explanee, and argued for the need to model explanation as an interactive conversation instead of a static object [Miller et al., 2017; Anjomshoae et al., 2019; Mittelstadt et al., 2019]. “Explanation naturally occurs as a continuous interaction which gives the interacting party the ability to question and interrogate the explanations.” [Madumal et al., 2018]. However, most contributions to explainable AI do not yet take this aspect into account [Henin and Métayer, 2019].

For explainable search, maybe more than for other subfields of explainable AI, such interactivity is a must: Not only is it unlikely that a single explanation of a search can answer all questions of a user (“You say that I should make this
Amir et al. and did not have a good idea of what strong policies look like, the explaining agent. If the human was not more competent human explainee is generally more competent at the task than the opposite assumption seems to be made: namely that the user’s misconceptions” [Sreedharan et al., 2020], which reduces the purpose of explanations to e.g. “correcting user’s misconceptions” [Sreedharan et al., 2020a] or bringing “the human’s mental model closer to the robot’s estimation of the ground truth” [Sreedharan et al., 2018], eventually serving “to convince end users to implement the recommended actions” [Dodson et al., 2011]. Even when explanations around the potential need for re-planning are theoretically discussed, the agent is expected to know best when to re-plan and when not to re-plan [Fox et al., 2017]. The same assumption on one-sided explanations is also often made in work on explainability in RL: Agent policies are commonly assumed to be optimal [Elizalde et al., 2008; Khan et al., 2008; Dodson et al., 2011; Khan et al., 2009; Tabrez and Hayes, 2019], and explanations are deemed necessary for example because “the limited human planning horizon and human spatial efficiency can greatly affect task performance” [Lee et al., 2019], or “a human’s sub-optimal decision-making could be attributable to a malformed policy given an incorrect task model” [Tabrez and Hayes, 2019].

In some research on summarizing entire agent policies, on the other hand [Huang et al., 2018; Amir and Amir, 2018; Amir et al., 2018; Lage et al., 2019; Sequeira et al., 2019], the opposite assumption seems to be made: namely that the human explainee is generally more competent at the task than the explaining agent. If the human was not more competent and did not have a good idea of what strong policies look like, how could she judge an agent’s strengths, weaknesses, and general trustworthiness from a number of example behaviors in “critical states” or otherwise “typical” situations? Despite the seemingly opposite basic assumptions on relative skill, the proposed explanations are still a one-way street however, with information only flowing from the agent to the human.

In explainable search, facing a typical online setting of sequential decision-making, solving any given task to completion or to optimality is rarely possible, and thus neither the searching agent nor the human interacting with it can generally be assumed to be omniscient. While a beginning student of chess, for example, will mostly rely on the stronger chess skill of an AI tutor, a grandmaster using the same AI for training might occasionally have good reasons to disagree with a particular analysis, and might have relevant aspects to add that the AI neglected; and the human supervisor in the search & rescue scenario will have access to different streams of information, sensor readings, and mission updates than the supervised robots, implying complementary roles in the overall task. This means that implicitly (through behavior in the task at hand) and/or explicitly (through communicative actions), explanatory knowledge has to flow both ways. While previous work assumed the ground truth to be on either the agent or the human side, explainable search can be thought of as a joint search for the ground truth: The best course of action cannot always be found by the human or the AI agent alone. In such settings of constant planning and re-planning, conversation [Hilton, 1991], argumentation [Zeng et al., 2018], contestability [Mulligan et al., 2019], and collaboration [Epstein, 2015] are key concepts for explainability. Explainable search is the ideal application for these broader challenges, potentially enabling “explanations to respond to the expertise and other context-specific needs of the user, yielding decisions that leverage, and iteratively learn from, the situated knowledge and professional expertise of users” [Mulligan et al., 2019]. In the ideal case of human-AI collaboration through explainable search, “on the one hand, explanations improve the cooperation, and, on the other hand, cooperation permits to each agent to produce relevant explanations for the other” [Brézillon and Pomerol, 1997].

Explanations as a two-way street. Previous work on explainable agent in classical planning, the optimality of the agent’s plan is generally assumed [Sreedharan et al., 2018; Sreedharan et al., 2020a; Chakraborti et al., 2019a], which is the ideal application for these broader challenges, potentially enabling “explanations to respond to the expertise and other context-specific needs of the user, yielding decisions that leverage, and iteratively learn from, the situated knowledge and professional expertise of users” [Mulligan et al., 2019]. In the ideal case of human-AI collaboration through explainable search, “on the one hand, explanations improve the cooperation, and, on the other hand, cooperation permits to each agent to produce relevant explanations for the other” [Brézillon and Pomerol, 1997].

Explanations in long-term interactions with users. It has long been understood that “to collaborate effectively with a person, a collaborative intelligence must be able to model the human view of the world.” [Epstein, 2015] This is also relevant for explainable AI, since “explanation naturally (... involves two processes: a cognitive process and a social process. Most prior work is focused on providing explanations without sufficient attention to the needs of the explainee, which reduces the usefulness of the explanation to the end-user.” [Madumal et al., 2018] Research into user-aware, personalized explanations is still relatively uncommon [Anjomshoae et al., 2019], even if there is interesting work in the classical planning setting on framing explanation as the reconciliation of the agent’s and the user’s model of a given task [Chakraborti et al., 2020; Chakraborti et al., 2019a].

For explainable search, challenges around user-awareness are front and center due to the envisioned long-term interactions with users – both within single sequential decision-making tasks or episodes, as well as potentially over many
such tasks during the lifetime of the agent. Long-term inter-
actions for example mean that the user’s understanding of the
world cannot be assumed to be given and unchanging as in
the majority of user-focused related work [Chakraborti et al.,
2017], but has to be inferred online based on prior knowledge
of the user, on estimating the influence of newly incoming
information on the user, and on direct questions to the user.
Repeated interactions with the same user open research ques-
tions around the learning of user preferences and satisfaction
models over time, and also allow for explanations to aim at
long-term user satisfaction [Kraus et al., 2019].

Explainable search exposes the research gap that “all of the
work on the topic of interpretable behavior has, unfortunately,
revolved around single, and one-off, interactions and little at-
tention has been given to impact of evolving expectations in
longer term interactions” [Chakraborti et al., 2019a]. Imag-
ine for example the limited use of a chess tutor agent that is
only able to explain individual moves, rather than accompany
users through entire games, and entire lessons consisting of
multiple games, while tracking their learning progress and
understanding. As a positive example, consider a search &
rescue robot that does not repeat certain elements of explana-
tions over and over again, because it is aware that the human
supervisor has already understood them when discussing a
previous decision just minutes ago, and knows that only con-
veying additional information is of value; moreover, imagine
the robot remembering from earlier interactions – maybe even
from past missions – the working style of the supervisor, and
what kind of information at what level of detail she prefers
when asking for explanations, so as not to put too many cog-
nitive demands on her while supervising multiple robots si-
multaneously. We believe that explainable search can make
ideal use of its online setting to tackle the challenge that “in-
telligent agents and humans need to be able to mutually ex-
plain to each other what is happening (shared awareness),
what they want to achieve (shared goals), and what collabor-
ative ways they see of achieving their goals (shared plans
and strategies)” [van Harmelen, 2020]. Through such adap-
tation over time, explainable search agents should be uniquely
suited to becoming a valued and trusted partner.

**Explanation-aware search.** The majority of works on ex-
plainable autonomous agents so far has considered explana-
tions “after the fact”, i.e., something that happens, possibly
in response to user questions, after the decision-making
process is finished. Not only is this insufficient in the light
of explanation as a two-way street, as discussed before, but it
also removes the opportunity of folding “the possibility of
having to explain its decisions (...) into an agent’s reasoning
stage itself” [Chakraborti et al., 2020]. In some recent work –
in the explainable classical planning setting [Chakraborti et al.,
2019b; Sreedharan et al., 2020a], as well as in a rein-
fforcement learning setting with users who have an incomplete
understanding of the rewards [Tabrez and Hayes, 2019] – the
idea was developed to treat explanations as “explanatory ac-
tions” instead. Added to the traditional “task-progressing ac-
tions”, these are actions with epistemic effects, actions that
can affect the user’s understanding of the task. Provided a
definition of suitable epistemic goals in additional to the tra-
ditional goals in the task at hand, these could be fully inte-
grated into the search process of an explainable search agent,
in order to explore context-specific and user-specific trade-
offs between task performance and explanation performance.

Imagine for example a search & rescue robot that is able
to proactively avoid surprises on the side of its human super-
visor by using behaviors it expects to be more easily under-
standable, for example when the supervisor is currently too
busy with other tasks (see also [Gervasio et al., 2018]). In
other situations, when the robot knows that either itself or
the supervisor is missing multiple pieces of information to
fully collaborate in the ongoing mission, it could trade off the
amount of explanatory communication with its degree of per-
formance in the task, or make the supervisor aware of such
trade-offs e.g. due to limited time. In this case, joint perfor-
mance could increase by avoiding an explanation Bottleneck.

Explainable search also opens up more research potential
in the adversarial setting of acting while hiding, instead of
explaining, plans or goals of the agent – a setting recently de-
veloped for the classical planning case as well [Keren et al.,
2016; Kulkarni et al., 2019]. Our chess tutor example even
yields possible non-adversarial applications for such obfus-
cation: The goal of an ideal teacher in any given lesson could
be aiming in between full (collaborative) explanation and full
(adversarial) obsfuscation of the teacher’s behavior. The goal
could be producing interesting decisions that introduce just
as much complexity to the student as that student is currently
able to handle – and ideally giving the student just enough
hints that she can figure out the core of the lesson by herself.
This kind of planning ahead would go far beyond current ap-
proaches of simply playing as well as possible, and only re-
acting to the user’s need for explanations when prodded.

**Counterfactual explanations of search.** Successful
search algorithms, such as those in the family of Monte
Carlo Tree Search approaches [Coulom, 2006; Kocsis and
Szepesvári, 2006], often handle large search spaces by
searching very selectively – by focusing on the most promis-
ding decisions and most likely states, and exploring the tree
mostly in their direction. This could lead to the following
challenge: What if the best explanation for a given behav-
ior is grounded in actions or states that were never explored,
or only explored very little, because the search was intelli-
gent enough to know that they would not matter for finding
a strong policy? Important reasons for a specific plan may
sometimes appear in the search that led to it, but they may
also sometimes be carefully avoided by that search.

Imagine a chess student asking, “Why is your move sug-
gestion so timid in this position?” It could be that the position
at hand allows for the user to boldly capture the enemy queen,
but this greedy move would lead to such an obviously unfa-
vorable exchange that the search, guided by advanced chess
knowledge, completely avoided wasting time exploring this
option. This example demonstrates that explainable search
to almost more than just explaining a search that just hap-
pened, for example by summarizing or visualizing the already
searched space. It also means interactively constructing the
searches that could have happened, and contrasting them to
the original choices the search algorithm made. In our ex-
ample, the results of the previous move search could for ex-
ample be compared to an additional search with stronger fo-
In a non-search context, constructed and used for contrastive explanations in prior work, the strategy itself may have still been well motivated and potentially optimal, and the search agent ultimately agrees with the user’s instincts. An important caveat is careful treatment of outcome bias [Baron and Hershey, 1988] and hindsight bias [Fischhoff and Beyth, 1975], since even if the user is right, the search strategy itself may have still been well motivated and potentially optimal in expectation under limited resources.

**Integrated explanations of search and evaluation.** According to dual-process theory [Kahneman, 2003], humans make use of two different modes of thought: System 1 refers to thinking that is fast, automatic, heuristic, and unconscious, such as for example visually recognizing an object. System 2 refers to thinking that is slow, effortful, calculating, and conscious, such as for example solving an algebraic equation. Connections have been drawn between this theory and some of the most notable recent examples of search-based AIs [Silver et al., 2017a; Silver et al., 2017b; Schrittwieser et al., 2019], as they consist both of a search algorithm which explicitly generates possible futures (System 2), and deep neural networks which heuristically evaluate these futures and guide the overall search process (System 1) [Anthony et al., 2017].

The aspects of explainable search we have outlined so far are concerned with explaining plans and action decisions via exploring the possible future scenarios that were generated during search (System 2). However, in order to arrive at a decision, these possible futures have to be evaluated and compared as well, for which a neural network might be used (System 1). Neural networks might also be guiding the search towards heuristically preferred action choices; in some cases they can represent other types of knowledge as well, such as the search algorithm’s understanding of its environment [Schrittwieser et al., 2019]. It is therefore a natural aspiration to ultimately combine our growing understanding of how to explain search with our growing understanding of how to explain neural networks. Only by integrating research on explaining data-driven systems into research on explaining goal-driven systems can we meaningfully illustrate and explain state-of-the-art search agents.

As one example for the resulting hybrid or holistic explainable searches, imagine an agent able to draw attention to surprises found during the search: states or actions that were initially believed to have low (or high) value – accompanied by an illustration of why the respective neural network gave this heuristic evaluation – but which in-depth search eventually found to be optimal (or disastrous) in the specific situation at hand – accompanied by an explanation of what makes this situation special. A related view of the two types of explanations involved here might be the distinction between process accounts (which “address the detailed decisions made during heuristic search”) and preference accounts (which “clarify the ordering of alternatives independent of how they were generated”) [Langley, 2019]. We see this distinction as complementary rather than identical to the distinction between algorithm-focused and domain-focused explanations – both a search tree (representing a process) and a neural network (encoding preferences) can for example be explained in terms of their internal processing, as well as in terms of what they represent for the application domain at hand.

**4 Conclusions**

In this paper we presented the challenge of explainable search. Search algorithms such as Monte Carlo Tree Search are used for planning, scheduling, decision making and optimization in countless research and application domains: from manufacturing to finance, from logistics and transportation to hospital planning, from software engineering to security modelling, from vehicle routing to materials design and discovery, from playing games to steering self-driving cars to acting on energy markets. Explainable search is therefore not only an interesting research challenge, but also potentially of great practical and economic value.

Explainable search will share advances with other subfields of explainable AI on questions such as how to develop domain-independent explanation techniques, how to best use theory of mind for generating explanations, or how to integrate verbal and non-verbal modalities of presenting explanations [Sado et al., 2020]; how to produce explanations in real time, how to best model user preferences, or how to explain complex environments with many interacting agents [Kraus et al., 2019]. At the same time, we believe that explainable search holds unique challenges, and promises unique gains, some of which we outlined in this paper.

In summary, we therefore propose to work towards the challenge of search-based agents that are able to explain their short-term decisions and long-term plans by explicitly reasoning about the complex spaces of possible futures which are spanned by their actions; agents that are able to do so in an online fashion in environments requiring sequential decision-making, continuously learning and adapting to the task and given resource limitations, as well as to the human user, their understanding, their needs, and their satisfaction with the task at hand; agents that are able to provide integrated explanations through methods designed to shine a light on modern machine learning as well as on state-of-the-art search; and agents that are continuously planning, communicating, and effectively collaborating with users through mutually understandable behavior. Through progress on the individual research questions outlined above, we envision such explainable search agents to become invaluable partners for human-machine collaboration, “allowing each (...) to operate in modes that utilize the strengths of both” [Crowder and Carbone, 2017], and establishing “a synergy between people and computers to accomplish human goals” [Epstein, 2015].
References


[Ehsan et al., 2019] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: a technique for explainable


[Kraus et al., 2019] Sarit Kraus, Amos Azaria, Jelena Fiosina, Maike Greve, Noam Hazon, Lutz M. Kolbe, Tim-Benjamin Lembcke, Jörg P. Müller, Sören Schleibaum,


