TRUSTWORTHY HUMAN-AI PARTNERSHIPS

SARVAPALI D. RAMCHURN, SEBASTIAN STEIN AND NICHOLAS R. JENNINGS

Summary

In this paper, we foreground some of the key research challenges that arise in the design of trustworthy human-AI partnerships. In particular, we focus on the challenges in designing human-AI partnerships that need to be addressed to help humans and organisations trust their machine counterparts individually or as a collective (e.g., as robot teams or groups of software agents). We also aim to identify the risks associated with human-AI partnerships and therefore determine the associated measures to mitigate these risks. By so doing, we will trigger new avenues of research that will address the key barriers to the adoption of AI-based systems more widely in our daily lives and in industry.

1. Introduction

Recent advances in Artificial Intelligence (AI), Machine Learning (ML), and Robotics have significantly enhanced the capabilities of machines. Machine intelligence is now able to support human decision making, augment human capabilities, and, in some cases, take over control from humans and act fully autonomously. Real-world examples of autonomous systems including self-driving vehicles, recommender systems, facial recognition systems, and automated trading are only just beginning to demonstrate the value machine intelligence can deliver to society and the wider economy. However, unfortunately, there have also been failures in the deployment of such autonomous systems that have resulted in fatal car crashes, plane accidents, and stock market failures (Brynjolfsson and McAfee, 2014; Daugherty and Wilson, 2018). Such failures are often attributed to a poor understanding of how to weave AI systems into our societal and industrial fabric.

Given this, we believe that the next big advance for AI and ML systems will involve them being significantly more tightly embedded into systems alongside humans, interacting and influencing each other in a number of ways. Such human-AI partnerships are a new form of socio-technical system in which the potential synergies between humans and machines are much more fully utilised. To achieve this, AI systems will need to leave their currently solipsistic nature behind and be able to *cooperate, coordinate, and compete* with one another and their human interlocutors. Such partnership will combine their complementary skills and capabilities to make the best use of the distinctive strengths of humans and machines (Licklider, 1960), while also acknowledging their potentially diverging preferences, purpose, and objectives that may give rise to conflict or cause them to attempt to influence each

The lead contact is Sarvapali D. Ramchurn (sdr1@ecs.soton.ac.uk).

other's decision-making (intentionally or not).[1] Likewise, humans will likely be challenged to work and live with AI systems as fully autonomous partners, rather than purely tools that they can manipulate or query. The modalities through which they engage with AI systems will also vary greatly, shifting away from typical screen-based or tactile interfaces to voice or brain-controlled, opening new opportunities and risks for interactions between humans and AI systems.

Designing, building, and deploying human-AI partnerships present a number of new challenges as we begin to understand their impact on our physical and mental well-being, our personal freedoms, and those of the wider society. Indeed, the deployment of machine intelligence within systems that are traditionally human-driven requires careful consideration of not only the reliability of these machines to take over human decision-making tasks, but also the ethical, legal, and psycho-social implications of the use of such machines. To tackle these issues, research communities, such as the Safe AI or AI Ethics communities, have rapidly grown,[2] and new international initiatives have emerged in the last few years, such as the Leverhulme Centre for Future intelligence,[3] the NYU Centre for Responsible AI,[4] Stanford University's Human Centred AI institute,[5] or the UKRI Trustworthy Autonomous Systems Hub.[6] From a technical perspective, *human-centred* (Lepri et al., 2021; Shneiderman, 2020; Wilson and Daugherty, 2018) and *machine-centred* approaches (Rahwan et al., 2019; Awad et al., 2018; Kraus et al., 2020) to the development of artificial intelligence have emerged. Human-centred approaches propose to develop socially beneficial and ethical machine intelligence that also augments human capabilities and undertakes human tasks with high reliability. The former seek to ensure humans are in control or at least have a meaningful oversight of machine decisions, helping to foreground machine decisions in ways that humans can comprehend and manage through various modalities. In contrast, machine-centred approaches aim to internalise human values because, in many situations, humans may not be able to intervene or have the ability or time to fully understand what the machine is about to do. While both approaches help focus research questions on the technical challenges involved in making machine intelligence reliable and augmentative, they represent different points on the spectrum of human-AI partnerships and raise key interactional and ethical challenges. It is also important, however, to consider the psychosocial aspects of systems involving humans and machines (i.e., the interdependence between social factors and individual behaviours and thought and decision processes). Indeed, in previous work (Jennings et al., 2014), we started to capture such aspects within the framework of Human-Agent Collectives (HACs). HACs are systems where neither humans nor

[1]These elements of the human-AI partnership build upon and generalise notions of human-machine interaction or human-machine teaming that largely focus on the cooperative and coordination aspects.

[2]These include venues such the AAAI/ACM AI Ethics conference (`https://www.aies-conference.com`), the AI Safety workshop at the International Joint Conference on AI (IJCAI) and the SafeAI workshop at the AAAI Conference on AI.

[3]`http://lcfi.ac.uk/.`

[4]`https://engineering.nyu.edu/research-innovation/centers/center-responsible-ai/.`

[5]`https://hai.stanford.edu/.`

[6]`http://www.tas.ac.uk/.`

machines are always in charge and there is a need to interact in various organisational and interactional setups in order to achieve their individual and common objectives. The HAC framework also accounts for the fact that humans and machines may need to be incentivised to undertake tasks, and their actions need to be traceable to ensure that the system is accountable.

The HAC framework, nevertheless, has its own limitations. For example, it does not consider the ethical or regulatory issues in diminishing human autonomy or tracking human activity, nor does it explicitly consider how human or machine failures are managed within such systems. The challenge is exacerbated by the fact that the AI and ML systems may involve networks of black-box autonomous systems that make it hard to trace the causes and the widespread consequences of individual failures. Indeed, as with much research within the AI community, projects tend to focus on efficiency rather than resilience, acceptance, or other intrinsic or extrinsic performance measures as viewed by different stakeholders. Furthermore, work at the boundary between AI and HCI (e.g., human-agent interaction or human-machine teaming) also tends to ignore the competing interests of the many stakeholders that may be involved and the strategic decisions they need to make. The result is that systems are built for lab-specific contexts and potentially never make it to real-world applications. In cases where such systems are deployed in particularly dynamic and uncertain environments, they will have to deal with events that neither the machines nor the humans were prepared to deal with. Such concern is supported by recent events. For example, recent self-driving car accidents involved the driver placing too much confidence in the capability of the autonomous car (e.g., either sleeping or watching videos, meaning they had no view of emergency alerts) (Banks et al., 2018), while the Air France 447 crash exemplifies how a combination of system failures, the lack of team coordination and human skill in making sense of data generated by interconnected sensors, and in manoeuvring with no automated help, led to catastrophic consequences (Salmon et al., 2016). Furthermore, our reliance on machines to comprehend masses of data from a variety of interconnected sources (e.g., sensors, databases, and streaming broadcasts) in real time, in order to make decisions on our behalf, means that we are exposed to the flaws built into these algorithms and the data they are trained on. In this vein, cases of bias in datasets and algorithms have led to serious discrimination in the selection of candidates for jobs or racially biased predictive policing (Daugherty and Wilson, 2018).

This has led to questions as to whether human-AI partnerships are feasible and financially viable as they may lead to costly consequences or require more expensive safeguards and monitoring than non AI-powered systems (Weardale, 2020). In some cases, developers of such systems have simplified the interactions between humans and machines, always leaving humans in control (e.g., voice assistants asking for confirmation before making automated purchasing decisions or reducing the autonomy levels in self-driving cars). Some have even recommended that black-box approaches should be avoided unless they can explain themselves (Office for Artificial Intelligence, 2020; National Security Commission on AI, 2020). We argue here that even if individual autonomous behaviours can be simplified or disambiguated, the interconnectedness of heterogeneous autonomous machines and humans (each with their own preferences, values, and interests), will create emergent system

behaviours that remain difficult to measure, predict, control, or certify. This could, in turn, undermine users' trust in these systems because of the perceived risks they pose. New approaches are therefore needed to design and oversee human-AI partnerships. We strongly advocate Responsible Research and Innovation (involving reflection, responsiveness, and anticipation as key elements) in the development of all human-AI partnerships to ensure they are ethical and safe. We believe the challenges presented in this paper will be helpful in discussions with a wider set of stakeholders beyond the research community. At a societal level, new approaches are required for the governance and regulation of such partnerships to ensure the actors involved remain accountable and industry is provided with clear guidelines on how these systems should be developed. At a technical level, a range of computational and interactional issues require urgent attention. For example, new models and interaction mechanisms need to account for the fact that machines will be able to continually learn and adapt in complex environments, and where data generated by humans and sensors may be liable to bias and errors. Such approaches will also need to deal with actors that may have their own competing interests, acting rationally to maximise their individual utility.

In this paper, we will foreground some of the key research challenges that arise in the design of trustworthy human-AI partnerships.[7] We define human-AI partnerships to be *trustworthy by design* if the humans and machines involved can rely on each other, self-organise to take advantage of each others' strengths and mitigate their weaknesses, and can be held accountable for their actions. In particular, we focus on the challenges in designing human-AI partnerships that need to be addressed to help humans and organisations trust their machine counterparts individually or as a collective (e.g., as robot teams or groups of software agents). We also aim to identify the risks associated with human-AI partnerships and therefore determine the associated measures to mitigate these risks. By so doing, we will trigger new avenues of research that will address the key barriers to the adoption of AI-based systems more widely in our daily lives and in industry.

The rest of this paper is structured as follows. In Section 2 we elaborate on the issues in developing and deploying continually learning systems that adapt to human needs. Section 3 details the main challenges in designing the interactions between humans and machines when they work together as part of teams. We also discuss how such teams can be verified to be trustworthy at design time. Section 4 elaborates on incentive engineering challenges, articulating the need to model human preferences and designing incentives to engage them in social welfare maximising behaviours. Section 5 provides two user stories that bring together the range of challenges explored in realistic ways. Finally, Section 6 concludes.

[7]We will also touch upon some of the ethics, governance, and regulation challenges, but a detailed exposition of such issues, in specific application domains, is well covered by the Safe AI and AI Ethics communities.

2. Trusting Data-Driven Human-AI Partnerships

As with all data-driven technologies, the performance of machine learning systems can be significantly affected by the quality of the data they are trained on, both in terms of the distribution of the data and the methodology used to collect the data. In this context, a particular concern is algorithmic bias (Danks and London, 2017). Here, we focus specifically on how new forms of bias may arise when human-AI partnerships are not just deployed in one-shot applications, but in long-lived systems where both humans and machines continually learn and adapt their behaviours.

2.1. **Positive Feedback Loops.** The obvious cases of algorithmic bias can be seen as reflections of biases that exist in our processes, organisation design, and other social constructs. As discussed by Perez (2019), even infrastructure, vehicle design and academic career progression are all biased at the outset against women. Babuta and Oswald (2019) show that issues of bias in predictive policing result in discrimination towards ethnic minorities. Moreover, Pink (2019) demonstrate that time-based biases impact the outcome of court judgements, the performance of medical staff, or the performance of children at mathematics tests in their early years. Hence, it is no surprise that machine learning systems that are driven by human-generated data or human-designed organisations may be liable to perpetuate significant degrees of bias. That is not to say that such biases cannot be identified and dealt with at source, but in many cases, they cannot be detected until the machine learning systems are tested on adversarial cases that are a poor fit to the training dataset. For example, Slack et al. (2020) even show that standard bias-detection (through explanation) tools can easily be fooled. More worryingly, the emergence of new biases is typically ignored. Techniques such as saliency maps and attention networks can help both explain how machine learning systems place more weight on specific features in the data and identify biases (Gilpin et al., 2018). However, weeding out biases in an online, continually learning system, remains a challenge.

As humans and machines establish long-lived partnerships, the data that machines consume to continually retrain themselves may come from the very humans their outputs are meant to influence. This is the case, for example, where a user attempts to teach a smart thermostat the temperature profile they prefer. If on some day, it feels too hot, the user will again have to change the setting and the machine will learn again to set the temperature accordingly. In some cases, the rate of change in user behaviours or preferences may mean that the machine continuously has to play catch up with the human. The thermostat may be constantly trying to satisfy the human, and when it fails to do so, gets turned off. In other cases, pernicious *two-way biases* (i.e., both human and machine influence each other's biases) are introduced by the partnership and these are hard to weed out. For example, a system where a CCTV operator uses a machine learning system to classify people as potential intruders may train the system to recognise people of a certain age, skin colour or wearing specific types of clothes based on their own subjective views. The authority relationship between the human operator and the machine learning system implies that the learning system should accept these new inputs as ground truth. Thus, a positive feedback loop is established, whereby the CCTV operator will reinforce their personal biases, while

doing the same for the learning system. Similarly, autonomous vehicles trained on driver behaviours[8] where such drivers may be aggressive, will bias the system to drive aggressively, in turn, reinforcing this behaviour in the driver and others using a similarly trained system. Avoiding positive feedback loops is crucial, if these systems are to be embedded in our daily lives in the long term, but detecting that such loops exist in the first place and how to address them remain a challenge. It may be possible to monitor and audit for individual user bias through independent review processes, but these can be costly and time-consuming and therefore cost-effective solutions are needed. At the same time, ongoing discussions with domain experts, individuals affected by automated decisions and the general public are needed, in order to establish appropriate value systems that human-AI partnerships should operate in (Friedler et al., 2021).

It may also be necessary to design AI to push back and police the humans' flawed behaviours. This is even more important if humans can maliciously game the machines, either to avoid detection or to benefit from their poor performance. For example, during the Ushahidi crowdsourcing effort in Haiti, in some cases, members of the crowd lied about their state (e.g., number of people in a settlement or their condition) to influence the deployment of resources (Norheim-Hagtun and Meier, 2010; Heinzelman and Waters, 2010). Microsoft's Twitter bot was maliciously trained with racist and rude language by Twitter followers for pure fun (Neff and Nagy, 2016). To address this challenge, greater emphasis needs to be placed on policing machine learning applications to detect attacks or verify human inputs. For example, Simpson and Roberts (2015) show how poorly trained human labellers can be detected by the machine learning algorithm and then presented with new tasks to upskill them. However, this may be difficult if the majority of inputs ingested by the machine comes from malicious sources. New techniques are needed to address this. Crucially, such techniques need to ensure that the costs of the associated remedial measures do not outweigh the benefits of the system.

Finally, even if some of these measures are put in place, the positive feedback loops induced will make it increasingly difficult to trace the root cause of the degradation of continual learning systems. Are these failures due to a delay in adaptation or due to sensor failures or both? This problem is exacerbated in interconnected machine learning systems whereby one machine learning system's output (e.g., faces classified as potential criminals or stock price predictions) feeds into another system as an input (e.g., to decide whether someone should be monitored by police or to drive trading decisions of many algorithmic platforms). For example, during stock market shocks, algorithmic trading systems may react to market recovery having learnt to behave in the same way as millions of other similarly trained algorithmic trading platforms. The market may then suffer significant oscillations that are difficult to manage and stabilise. As such systems fail, it may be too late for humans to trace the causes and devise an alternative plan. In recent work, Pearl (2019) outlines a number of techniques for causal reasoning that may be able to learn causal relationships in data and traces. Furthermore, the author proposes the algorithmisation of counterfactuals to weed out causes of certain biases or behaviours. These are promising

[8] https://towardsdatascience.com/teslas-deep-learning-at-scale-7eed85b235d3.

avenues of research. However, additional work is requiredy to apply to highly unstructured domains where relationships can be highly convoluted (as in the case of positive feedback loops involving multiple actors). In particular, novel approaches are now needed to measure the impact and range of potential human and machine failures and new approaches designed to verify these systems at design time. We expand on this issue next.

2.2. **Tracking Data and Decisions.** In human-AI partnerships, data can be generated through interleaved and sometimes joint actions. In many cases, only the final outputs may be visible, while, in fact, at every point in the data pipeline, biases and errors are introduced. For example, in a disaster response scenario (Ramchurn et al., 2015), first, aerial imagery may be produced by UAVs. Second, humans and computer vision algorithms may work together to classify the images. Third, another set of humans may extract key targets for food delivery, and collaboratively optimise the allocation of these targets with planning agents and another set of UAVs autonomously flying in the air space. The UAVs, in turn, may need to iterate their routes with their human operators to maximise team performance and achieve their goals efficiently. In this process, humans and machines collaborate on tasks, jointly generating labels for images, interleave their decisions (i.e., humans and machines create tasks for UAVs, UAVs generate routes collaboratively with humans), and then humans may take control of the machines when more granular control is needed (i.e., route optimisation). Here, the failure of any human or machine could cause either a lost food package or an accident. Due to the interdependencies between humans and machines in the decision chain, attributing liability to either human or machine (or both) can be a computationally (and legally) challenging task. In some cases, the various relationships between the humans and the machines may already be known and can be modelled using techniques such as Probabilistic Reasoning or Bayesian networks. As shown by Venanzi et al. (2014), Bayesian approaches can be effective in weeding out biases and unreliable actors, while still delivering high quality outputs. In other cases, where decisions are based on other humans and machines working together (e.g., jointly labelling a picture or creating a plan), attributing liability and detecting where inconsistencies arise in the decision making process may be difficult, unless all the data and decisions made by the human-AI system are tracked and monitored.

To provide an audit trail of such human-machine interactions, software engineering approaches such as provenance tracking have been proposed as a potential solution (Moreau et al., 2008). Provenance tracking solutions enable the capture of data and decisions in a machine readable and traceable format. These methods generate directed acyclic graphs that can be analysed for inconsistencies and to identify the source of errors. However, in large systems producing large amounts of data every second, this analysis rapidly unravels. This is even more difficult when data is held in proprietary databases. The challenge then is to develop decentralised solutions that allow for the independent or containerised verification of causes and effects (Chaudhry et al., 2015). Crucially, these techniques need to efficiently work on abstractions of histories or summarise and characterise the essential features of a provenance trace in a way that still guarantees some traceability to the root cause of specific outcomes. It is also important to design such provenance abstractions

that can be interrogated and validated to ensure they can be trusted by both humans and machines, for example using graph clustering or network analysis techniques (Huynh et al., 2018). Otherwise, they would merely add another level of opacity to the system and increase the workload of those managing the system.

3. Designing Human-AI Partnerships

In human-AI partnerships, collaboration typically involves humans either making decisions individually or as a collective, working closely with autonomous machines. While a wealth of research exists in the multi-agent systems literature to deal with machine to machine conflicts, for example using argumentation or negotiation (Rahwan et al., 2003; Baarslag et al., 2014; Vasconcelos et al., 2009), very little work has been done to address conflicts in human-AI partnerships, particularly those involving multiple humans and multiple machines, going beyond constrained negotiation scenarios (Lin and Kraus, 2010). Here, we discuss, with a practical example, the challenges that arise in such partnerships.

Building on the disaster response example above, a human-UAV team could involve UAVs being dispatched with an initial set of goals determined by their human team-mate(s) and the UAVs, in turn, reacting to their environment, changing their plan or suggesting to the human team on the ground to move to a different position. Humans may also suggest to the UAVs new goals or constraints on their actions, and negotiate with them over the viability of different courses of action. The UAVs may be expected to explain why they would refuse to take certain paths or estimate how achievable are the goals suggested by the operator (given the UAV's perceived risks in the environment). Another example may be a set of smart homes (each operated by a software agent) and human occupants, working together to adapt their electricity consumption profile in response to degraded supply on the power grid. The software agent may suggest to the occupants to run their washing machine at off-peak times to save money, while the human may indicate to the agent that they are unlikely to be home over the weekend to help the agent schedule heating and automated appliances optimally. In these different examples, one may have a human in charge ("on the loop") or engaged directly in the operation of the team ("in the loop"). Over time, these partnerships may also change from flat to hierarchical structures (where either the human or the machine is the leader) or teams may need to disband to create new ones (e.g., if a sub-team of UAVs needs to be formed for a specific task or smart homes need to work together as part of energy collectives to purchase energy contracts). Managing such relationships requires engineering the interactions, validating team structures, and ensuring that these teams remain accountable. We address these issues next.

3.1. **Designing Interactional Arrangements.**

Here we consider the problem of configuring the roles and relationships between humans and machines in teams. Roles tend to be allocated based on the core capability of each member of the team, while considering the requirements or constraints of the environments. Typically machines are given the most data intensive or repetitive tasks (e.g., monitoring the environment, calculating optimal routes or allocating tasks) while humans perform mostly reasoning tasks (e.g., setting goals

and reacting to unseen situations) (Tambe, 2011; Verame et al., 2016). In some cases, machines may be allowed to take over from humans in situations where they can be trusted to operate at a similar level and can always be turned off if they underperform (Verame et al., 2016; Costanza et al., 2014). In many cases, the starting point for the team formation process will be the existing human-only setup with machines as replacements (e.g., to control a vehicle or to extract insights from datasets) or enhancements (e.g., to reduce mental workload for a CCTV operator or a doctor).

Doing so, however, risks increasing workload for humans in dynamic situations and under-exploiting more efficient human-AI partnerships. As seen in work by in Ramchurn, Wu, et al. (2016), a team scheduling agent (replacing a team leader) that gives some control to multiple members of a human team to choose their preferred actions, can result in poor choices at the team level. This is because the machine is unable to understand human constraints and cannot predict and inform them of the consequences of their choices in many cases. As shown by Ramchurn, Huynh, Wu, et al. (2016), having a human mediating between the machine and the humans can achieve the best performance, also outperforming a purely human team. These *interactional arrangements* dictate the architecture of a human machine team and should take into account not only the capabilities of humans and machines, but also the respective requirements they place on each other. For example, machines may need one or more humans to interpret their outputs and humans need to be able to predict the reaction of machines to their inputs. Machines also need to be able to provide means for humans to interrogate them to better understand their suggestions or behaviours (as we discuss next). Finally, roles may need to be specifically created to analyse and manage human-AI interactions rather than purely focusing on task performance (Daugherty and Wilson, 2018).

Going further, it may be possible to configure the machine behaviours and controls to suit the specific humans they need to work with (e.g., the young, tech-savvy, or elderly). For example, machines may provide extra controls for expert users and adjust their level of autonomy according to the experience of the individual user or team interacting with it. Their autonomy levels could also be adjusted based on the configuration of the human team and the goals they would like to achieve. For example, a swarm of UAVs may decide to split into smaller swarms if the tasks at hand require granular control and where there are sufficient operators to manage their actions. The machines could either suggest the new plan to their tactical commanders in charge of planning the mission (and wait for approval or modifications) or could directly call upon the operators (in charge of low-level control) to join their team in case they are readily available and there is no time to overload their commanders with suggestions. In turn, commanders may want explanations to be provided by the machines to justify their choices and may decide the plan does not meet their objectives and override it. Similarly, in a smart home, a software agent could decide to switch the home to a better low-carbon energy tariff, and continuously watch out for better tariffs after having learnt the occupants' carbon/price trade-off. In case the tariff requires the home and its occupants to change their energy consumption habits (due to the intermittency of renewable power), the agent may create appliance and heating schedules proactively in order to nudge the users to fit the required profile. The humans could push

back on this with revised preferences or simply take complete control of the tariff switching process.

The adaptive behaviours and interfaces described above ineluctably introduce additional interactional challenges that could be tackled either by providing the AI system with a better understanding of human needs and states (e.g., using sensors or asking for input) or by designing the system to account for a range of human factors. For example, it is already possible to sense users' levels of stress (e.g., using EEG or ECG signals), frustration (from facial features) or tiredness (from their heart rate or time in working mode) in order to adapt their notifications and behaviours. However, the challenge is to determine what an AI systems should do if it detects potential stress or tiredness. While some have suggested using Bayesian models or reinforcement learning to learn the right action to take (e.g., Truong et al. (2016) model a user's bother cost or Ramchurn, Huynh, Wu, et al. (2016) model the preference of humans for certain team mates), such solutions remain difficult to scale up and validate in the real world where there is significant uncertainty as to the causes of stress, frustration, or tiredness. A possible solution is to centre the design process on the human-AI partnership, trialling different configurations of roles and AI capabilities in various iterations of the system in different settings, and iteratively improving the interactional modalities and machine intelligence at the same time. This may, for example, determine if the machines have the ability to deviate from plans (e.g., when the environment significantly changes or when the human's input is too delayed) and when to do so without human input (e.g., when the human is busy with a more important task or unable to communicate). In turn, humans may need to take on roles that are subordinate to machines in specific circumstances. Complex human-AI partnerships are likely to emerge in that process. However, the more complex such relationships become, the harder it is to verify and manage them, as we show next.

3.2. **Verifiable and Explainable Human-Machine Teaming.** Optimising the design of human-machine teams against specific performance metrics does not usually consider whether such performance can be guaranteed, nor be predictable in unseen contexts. Traditional methods to verify software typically involve knowing the answer to a given situation or knowing all the possible states in advance. Given a set of constraints or expected states, one can then test the system in a variety of settings to make sure that the system conforms to its intended behaviour. This approach can completely break down in human-machine teaming settings, particularly those that involve a high level of flexibility in terms of mixed-initiative decision making in dynamic and uncertain environments that can pose unpredictable challenges (Harel et al., 2020). Indeed, it is not clear how such a system can be verified to be compliant and safe if it can potentially change its behaviour (as in the case of machine learning systems discussed above) depending on new unseen inputs or if humans make the wrong decision when they are overloaded or tired. Traditional methods such as model checking or formal verification, would require modelling an infinite number of states to solve the problem, and also unrealistically assuming that human behaviour can be precisely modelled.

At best, autonomous systems tend to be developed, tested, and verified independently of the human behaviours or with assumed human responses. Recognising the limitations of this, new approaches have been developed to help system designers and users understand how human-AI systems reason or optimise their behaviours (Rodden et al., 2013; Gunning et al., 2019; Ribeiro et al., 2018). These approaches aim to make systems *interpretable* and *explainable*, in that they allow designers and users to interrogate and check what a system uses as key features based on its inputs (from humans or the environment), in order to generate an output (i.e., the link from the output to the most important parts of the inputs). These techniques also aim to provide an explanation to end-users that focuses on simplifying the complex elements of decision making and help them predict future behaviours of a system. For example, an image classifier may provide a means to identify key patient features or tissue features in an image that it uses to classify tissues as cancerous (Jansen et al., 2020). Another example may be an autonomous navigation system that provides visualisations of its environment using a point cloud to show where it thinks the obstacles are (Wu et al., 2018). However, most research in this space tends to focus on machine learning systems rather than other forms of AI, including human-machine teams (Samek et al., 2017).

Allowing designers to model and verify the potential behaviours of human-machine teams requires new tools to interpret the dependencies within such teams, simulate their behaviours efficiently, and calculate likelihoods of potential failures. Indeed, human-machine teams deployed in dynamic and uncertain environments essentially exhibit the properties of a complex system where emergent behaviours and the risk of cascading failures are vastly increased compared to traditional software systems where humans have complete control. Applying standard testing protocols (e.g., black box or glass box testing) may give false confidence in the certification of the system. Instead, we suggest the use of machine-generated explanations that consider both the fundamental operating principles underpinning a human-machine team (i.e., what dependencies, checks, and balances they contain) and a summary of their simulated behaviours or *digital twins* in a variety of setups (e.g., as evidence to support explanations or to demonstrate that how likely failures are and what mitigation measures they would adopt). These explanations present the user with the ability to run what-if questions to better understand the range of potential behaviours and their causes (Pearl, 2019). Moreover, when failures do occur, robust techniques are needed to model and ascribe responsibility within complex human-agent teams (Yazdanpanah et al., 2021). Going further, given the ethical issues that human-machine teams often raise (e.g., due to the harm they may expose humans or the infrastructure to), it is also important to take into account the societal acceptance, regulatory, and legal regimes within which they are implemented. This can only be achieved by involving experts from these disciplinary backgrounds in the design, implementation, and certification process. In turn, the cost of such processes may preclude small and medium sized organisations utilising human-AI partnerships from accessing such experts and iterating their product or service to meet their requirements. Hence, more work is needed to reduce such costs, potentially through the design of automated testing systems and standards that are easier to commoditise and scale up at low cost.

4. Incentive Engineering

Human-AI partnerships are rarely dyadic relationships, but rather exist in complex organisations involving many human and machine agents. These agents may have different, sometimes competing, aims and objectives. For example, they may be autonomous vehicles vying for space on an intersection or smart houses coordinating their electricity consumption within a neighbourhood to use a limited supply of renewable energy. To address these settings, we now elaborate on the socio-technical challenges of incentivising humans and machines to take actions that are not only beneficial to themselves, but also to the wider society. This is particularly difficult when the actors hold private information about their personal preferences and constraints and may not wish to share that information in case they are exploited. Hence, going beyond the notions of human-machine interactions, we next focus on the *modelling of preferences* and *engineering of incentives* to guide and sustain the formation of human-AI partnerships. On the one hand, models of preferences permit the design of machine behaviours and systems that better respond to humans, while the engineering of incentives ensures that human and machine behaviours are predictable and guarantee certain outcomes. To date, a large amount of work from the economics and AI literature has established a wealth of solutions and approaches to developing preference models and incentive mechanisms. However, they invariably make strong assumptions that need to be tackled if human-AI partnerships are to thrive, as we show next.

4.1. **Preference Modelling.**

To effectively assist or take decisions on behalf of humans, AI algorithms need to be aware of the preferences, needs and constraints of humans. This is particularly the case when making decisions that affect large groups of people, for example in smart electricity, smart transportation or disaster response applications. Here, a particular challenge is to collect preference information while safeguarding the privacy of users. One promising option for achieving this is to take a citizen-centric approach, where an individual is represented by a personal intelligent agent that learns and safeguards this information about private preferences and then interacts with AI systems on the user's behalf (Alan et al., 2016; Verame et al., 2018). As a concrete example, such an agent could observe the usual departure times and daily energy consumption of an electric vehicle driver and then use this information to participate in an electricity market or dynamic pricing regime. When the driver wishes to travel further than their vehicle's range, the same agent might also book a multi-modal journey using public transport, while taking into account the user's preferences for trading off speed and convenience with price. As such, an agent could run locally on a smart device or via a trusted third party, meaning no detailed private information has to be surrendered to a centralised decision-making system.

However, modelling the preferences of citizens accurately raises new research challenges. First, this has to be achieved without placing undue cognitive burden on users. A potential approach for doing this is to make use of inverse reinforcement learning to estimate a citizen's latent preferences from observations (Hadfield-Menell et al., 2016). This could be augmented with domain-specific preference and behaviour models, e.g., using mobility models (McInerney et al., 2013), thermal comfort models (Auffenberg et al., 2018) or discrete choice models (Västberg et al., 2020). To deal with uncertainty and unseen situations,

it will also be necessary to combine this approach with preference elicitation (Baarslag and Gerding, 2015), where the agent requests further input from the citizen if the expected benefit outweighs the cognitive cost. Finally, in order to address the cold start problem and further reduce costly user interactions, data could be shared between multiple intelligent agents to identify common patterns and provide suitable priors for the preference models, which are then refined through subsequent observations or queries. This sharing of data could be undertaken directly between trusted agents, e.g., between the agents of family members or friends, or it could be achieved via third-party preference aggregation agents that collect and anonymise data. Such approaches might build on existing techniques used in recommender systems (Aggarwal, 2016) or federated learning (Smith et al., 2017).

4.2. **Incentives.** When designing AI systems that interact with humans, it is critical to consider that these humans are not passive providers of data with fixed behaviour patterns. Rather, they are autonomous entities that pursue their own personal goals, and these may interact with the decisions of AI systems in unexpected ways. As argued earlier, this could lead to strategic manipulation, where a user purposefully supplies incorrect data in order to obtain a personal benefit. At the same time, there are also opportunities to align the objectives of an AI system closely with the goals of human users and proactively encourage positive behaviours, such as the reduction of energy consumption or delaying travel plans to reduce traffic congestion. To mitigate manipulation and encourage positive behaviour change, it is thus crucial to model the goals of users and provide them with the right incentives.

One prominent strand of research considers the use of game theory to achieve this and to model the behaviour of rational decision-makers in the presence of incentives (Nisan et al., 2007). This allows AI systems to incentivise truthful reporting of private data, thus minimising the potential for strategic manipulation (Albert et al., 2017). Related techniques from decision theory can be employed to model how users respond to varying prices or service quality levels, and this can be used to optimise overall system efficiency, for example to encourage participants in a car sharing system to help relocate vehicles to socially-beneficial locations (Drwal et al., 2017) or home occupants to adjust their thermostat settings to save on their energy bills, trading off their comfort (Shann et al., 2017). However, while much work in game theory and decision theory assumes rational behaviour, humans do not necessarily satisfy that assumption. For example, there is ample evidence that people do not evaluate all possible outcomes when making decisions, but rather choose solutions that are deemed good enough (Simon, 1955). For many decisions, the framing of choices is also important, leading people to choose different outcomes depending on how they are presented or what alternatives are available (Tversky and Kahneman, 1981).

Designing these incentive-aware systems is challenging, because accurate models about how (potentially irrational) individuals respond to incentives needs to be learnt. This is related to the modelling of preferences described earlier, but also needs to consider how individuals might strategically influence the learning process in order to derive an advantage in future interactions (Amin et al., 2013). More generally, when AI systems adapt their decision-making policies dynamically based on observations, this could allow

participants to exploit imperfect policies (Stein et al., 2020). In addition to addressing this strategic behaviour, it is also desirable to limit the amount of information required by a decision-making mechanism, in order to safeguard privacy.

5. User Stories

Here we present two user stories that aim to exemplify the challenges described in the previous sections, in different contexts and where the human-AI partnership elements perform well and also fail at times. We aim to show the broad range of issues that can arise but where there is commonality in terms of the fundamental research questions that need to be addressed. By so doing, we also aim to demonstrate the kind of impact that such human-AI partnerships can have if the challenges are met.

> **User story 1:** Sarah is woken up by her digital assistant, who she affectionately calls Alice, that runs on her alarm clock. Alice plays her preferred wake-up music and guides her through a mindfulness session to get her ready for her day. Alice asks Sarah for her top tasks of the day and helps her schedule them, avoiding conflicts with her other meetings. Alice can run from any of her personal devices, accessing Sarah's profile in the cloud at any time to provide her with a personalised experience in any context.
>
> Alice is particularly useful when planning Sarah's journeys to meetings with different clients across town. Each of her clients also use similar personal assistants but not all use Alice, mainly because the firm that runs Alice also shares their users' data with third party marketing companies and other subsidiaries of the firm.
>
> Today, Sarah plans to meet George over coffee to discuss a new property development project that just came up. To this end, Sarah asks Alice to negotiate a location and time with George's digital assistant called Bob. Bob is also run by a rival retail company to that of Alice.
>
> Alice requests Bob to share details of George's schedule for the day and whether George has any preference in terms of location or any constraints on travel. Bob refuses to share any details except George's availability at certain times due to the George's restricted data sharing agreement. Alice retrieves the history of interactions with George and determines that George typically likes to meet downtown. She proposes a list of coffee shops in that area to Bob, ranking the ones run by Alice's firm first. Bob accepts the top suggestion and the meeting is booked in George's calendar.
>
> Closer to the meeting time, George realises that the meeting has been set at a coffee shop owned by Alice's firm, which he dislikes. George calls Sarah and suggests a more preferred location and informs Bob never to book a meeting in that coffee shop again. Bob records the change of location and associates a negative preference with all coffee shops owned by Alice's firm.

This story reveals the issues involved in delivering AI-based services that model, learn, and adapt to one or more users' preferences and also autonomously negotiate on behalf of

their owners. The interactional arrangement (see Section 3.1) here involves agents acting as mediators for their human owners, but the relationship between human and agent is blurred by the firms that run Alice and Bob. Alice can continually learn Sarah's preferences and introduce biases as we see. It is unclear that Sarah is even aware that some of her choices may be shaped by the firm that runs Alice and it may be hard to determine who is eventually responsible for Sarah's choices. In the long run, Sarah's autonomy in making her own choices may be compromised (see Section 2.1) and Alice's biases could transfer to Sarah and vice versa (i.e., Sarah may only prefer to go to Alice's firm's retail outlets or Alice may learn to only propose specific coffee shops that Sarah chooses). Furthermore, while Alice can continually learn Sarah's preferences, it cannot easily do so with other human counterparts, though it attempts to restrict the choices (i.e., ranking the coffee shops) offered to Bob in an attempt to incentivise Bob to pick the top one (see Section 4.2). In George and Bob's relationship, it is obvious that Bob just cannot mediate on behalf of George effectively unless George consents to some extra information being passed on to third parties (see Section 3.1). The problem could be fixed by having Bob asking George to confirm the meeting in the first place. George is frustrated with the choice made by Bob (who is not at fault) and realises there is a gap in Bob's model. George eventually directly inputs his preferences to Bob (see Section 4.1). In this case, George's dislike of Alice's firm is strong enough to overcome the bother cost of specifying his preferences to Bob. This example highlights potential design choices at hand:[9] (i) letting agents learn user preferences; (ii) designing the agent to request confirmation before every automated decision; or (iii) asking the users to directly input their preferences. While (i) will meet typical machine learning challenges, (ii) and (iii) can frustrate users and eventually break the human-AI partnership. Finding the right balance will depend on a number of factors, including the users' mental model of the AI, the risks involved in getting choices wrong, and the benefit of machine-speed decision-making.

> **User story 2:** David and Tom are part of a disaster relief charity, working in a major disaster zone in an Italian city. An earthquake struck five hours ago, and David and Tom just landed at a nearby airport, tasked with undertaking a rapid recce (reconnaissance mission) to get an understanding of the status of road networks, building damage, and casualties. To help with this task, they have brought 3 containers carrying 1000 small unmanned aerial vehicles, each equipped with onboard computing and communication capability that allow them to deploy as a swarm and communicate live imagery in real time. As the lead Swarm Operator, David is in touch with his headquarters (HQ) based in France and 100 other similar teams, deployed within a 100km radius. Tom instead, is tasked with driving around the disaster area to verify information coming from the swarm and to secure food and water depots.

[9]There may be other design choices such as letting the user choose from a menu of options or asking the user to fill the gaps in the agent's knowledge when the need arises.

> At the press of a button on his tablet, David commands the swarm to take off, rapidly mapping the area. David sees the map coming alive with the location of blockages and damaged buildings, with associated likely casualties. David requests Tom to check on a supermarket in an industrial area to determine if it is safe to use it as a food distribution point and to secure it. Tom can ask for support from the swarm if needed. Three minutes later, a red light flashes on his screen telling him that a fire has erupted in the industrial area. The swarm has lost a third of its units in that area.
>
> David assumes the fire has taken them out. A few minutes later, Tom radios in to say he has reached the site and that all is clear and that he is heading back. David asks Tom if he requested the swarm to scout a dangerous area, which might have led to the swarm being degraded. Tom replies he did ask for scouting support but denies the crash is his fault.

This story highlights the benefits of reducing the complexity of managing a swarm of UAVs with a few humans. The highly decentralised decision-making structure (where both humans and machines are highly distributed) adds significant degrees of opacity when handling large-scale failures and dynamic environments (see Section 3.1). In such situations, it would seem larger teams requiring greater control and visibility of the swarm would perform better. The key challenge highlighted here is that of predicting the swarm's emergent behaviours and managing them. Tom may have requested the swarm to scout an area without quite understanding *how* the swarm would execute this. Ideally, the swarm would have predicted its own behaviour and explained potential risks to Tom based on where its members may end up and allowing Tom to verify if the risks were acceptable (see Section 3.2). Furthermore, the story suggests that the swarm had high levels of autonomy, making decisions to go into risky areas without David confirming it was allowed to do so (see Section 3.1). Ultimately, to confirm who is at fault for the losses suffered, the provenance of the UAVs' crash would need to be traced back to either David's or Tom's decisions (see Section 2.2). Assuring the data pipeline used by the swarm and the humans involved is key to ensuring the human-AI partnership is accountable for its actions.

6. Conclusions

We have outlined an agenda to address a range of high-level research questions to ensure that human-AI partnerships are trustworthy by design. While we have touched upon some of the ethical issues such as human autonomy in human-AI partnerships (e.g., in long-lived interactions or in large human-machine teams), we have not focused on the ethical, governance, and regulatory questions, as these are already under active investigation by a range of multi-disciplinary communities (e.g., AI Ethics, Human-Centred AI, and Safe AI). Instead, we explicitly focused on the technical challenges that need urgent attention and are currently underserved by the research community. To develop trustworthy human-AI partnerships, we discussed the need to consider: (i) the design of continually learning systems that work closely with humans and can create pernicious biases both for humans

and AI systems; (ii) the assurance of data pipelines used by human-AI partnerships and provided initial ideas on how accountability can be preserved in such systems using causal reasoning and provenance tracking; (iii) interactional arrangements to create effective organisational structures for human-AI partnerships; (iv) the verification and validation of human-AI partnerships through the creation of simulations of human-AI teams (e.g., using digital twins) and explanation tools; (v) the need to design incentives for human-AI partnerships to ensure they act individually and as collectives to achieve socially beneficial outcomes and (vi) the need to model the preferences of the actors involved to ensure fair and efficient outcomes for both humans and AI systems.

We presented two user stories to demonstrate how the issues of bias, accountability, control, and incentive engineering can all arise in both routine human-AI partnerships and highly dynamic and uncertain work environments where the human-AI partnerships demand high degrees of accountability. By going beyond settings involving one-human to one-agent, our stories serve to highlight specific issues with human-AI partnerships that are not the typical focus of existing fields of research (e.g., HCI, HAI, or AI Ethics). They also highlight that the development of trustworthy human-AI partnerships requires careful consideration of psycho-social, interactional, and engineering issues, as well as the preferences of the actors or the firms they represent.

Reflecting on the research directions proposed in this paper, we observe that they point to a stronger embedding of models of societal conventions and ethics into sophisticated (home and industrial) robots and software, interconnecting human and machine decision making. There is no question that the range and intensity of human-machine partnerships will grow significantly as AI systems become less solipsistic and more embedded in our daily lives, in industry, and even in government. Nevertheless, we believe the full benefits of human-AI partnerships will only be realised if they can be trusted to operate safely as they continually evolve. Such partnerships have the potential to improve our well-being, strengthen our society's resilience to natural and man-made disasters, and make our societies fairer and more economically efficient. This will require research into new AI technologies, as well as new methodologies for the design of human-AI partnerships, bringing together diverse disciplinary perspectives.

7. Acknowledgements

We would like to thank the reviewers for their very thoughtful comments, which led to a much improved version of the paper. This work was funded by AXA Research Fund, the EPSRC-funded Smart Cities Platform Grant (EP/P010164/1), and the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1). Sebastian Stein is also supported by a UKRI Turing AI Fellowship on Citizen-Centric AI Systems (EP/V022067/1).

8. Author Contributions

S.R. wrote Sections 1–4 and 6. S.S. wrote Section 5 and contributed to the other sections. N.J. contributed to all sections.

9. Declaration of Interests

The authors declare no competing interests.

References

Aggarwal, Charu C. (2016). *Recommender Systems*. Springer.

Alan, Alper T., Costanza, Enrico, Ramchurn, Sarvapali D., Fischer, Joel, Rodden, Tom, and Jennings, Nicholas R. (2016). "Tariff agent: interacting with a future smart energy system at home". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 23.4, pp. 1–28.

Albert, Michael, Conitzer, Vincent, and Stone, Peter (2017). "Automated Design of Robust Mechanisms". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI17. San Francisco, California, USA: AAAI Press, pp. 298–304.

Amin, Kareem, Rostamizadeh, Afshin, and Syed, Umar (2013). "Learning Prices for Repeated Auctions with Strategic Buyers". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS13. Lake Tahoe, Nevada: Curran Associates Inc., pp. 1169–1177.

Auffenberg, Frederik, Snow, Stephen, Stein, Sebastian, and Rogers, Alex (2018). "A comfort-based approach to smart heating and air conditioning". In: *ACM Transactions on Intelligent Systems and Technology* 9.3, pp. 1–20.

Awad, Edmond, Dsouza, Sohan, Kim, Richard, Schulz, Jonathan, Henrich, Joseph, Shariff, Azim, Bonnefon, Jean-François, and Rahwan, Iyad (2018). "The moral machine experiment". In: *Nature* 563.7729, pp. 59–64.

Baarslag, Tim and Gerding, Enrico H. (2015). "Optimal Incremental Preference Elicitation during Negotiation". In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. IJCAI15. Buenos Aires, Argentina, pp. 3–9.

Baarslag, Tim, Hindriks, Koen, and Jonker, Catholijn (2014). "Effective acceptance conditions in real-time automated negotiation". In: *Decision Support Systems* 60, pp. 68–77.

Babuta, Alexander and Oswald, Marion (2019). "Data analytics and algorithmic bias in policing". In: *The Royal United Services Institute for Defence and Security Studies*.

Banks, Victoria A., Plant, Katherine L., and Stanton, Neville A. (2018). "Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016". In: *Safety Science* 108, pp. 278–285. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2017.12.023.

Brynjolfsson, Erik and McAfee, Andrew (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Chaudhry, Amir, Crowcroft, Jonathon, Howard, Heidi, Madhavapeddy, Anil, Mortier, Richard, Haddadi, Hamed, and McAuley, Derek (2015). "Personal data: thinking inside the box". In: *Critical Alternatives* 1.1. ISSN: 2445-7221.

Costanza, Enrico, Fischer, Joel E., Colley, James A., Rodden, Tom, Ramchurn, Sarvapali D., and Jennings, Nicholas R. (2014). "Doing the laundry with agents: a field trial of a future smart energy system in the home". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 813–822.

Danks, David and London, Alex John (2017). "Algorithmic Bias in Autonomous Systems." In: *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4691–4697.

Daugherty, Paul R and Wilson, H James (2018). *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

Drwal, Maciej, Gerding, Enrico H., Stein, Sebastian, Hayakawa, Keiichiro, and Kitaoka, Hironobu (2017). "Adaptive pricing mechanisms for on-demand mobility". In: *AAMAS '17 Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. ACM, pp. 1017–1025.

Friedler, Sorelle A., Scheidegger, Carlos, and Venkatasubramanian, Suresh (2021). "The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making". In: *Communications of the ACM* 64.4, pp. 136–143. ISSN: 0001-0782. DOI: 10.1145/3433949.

Gilpin, Leilani H., Bau, David, Yuan, Ben Z., Bajwa, Ayesha, Specter, Michael, and Kagal, Lalana (2018). "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, pp. 80–89.

Gunning, David, Stefik, Mark, Choi, Jaesik, Miller, Timothy, Stumpf, Simone, and Yang, Guang-Zhong (2019). "XAI—Explainable artificial intelligence". In: *Science Robotics* 4.37. DOI: 10.1126/scirobotics.aay7120.

Hadfield-Menell, Dylan, Russell, Stuart J., Abbeel, Pieter, and Dragan, Anca (2016). "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems 29*, pp. 3909–3917.

Harel, David, Marron, Assaf, and Sifakis, Joseph (2020). "Autonomics: In search of a foundation for next-generation autonomous systems". In: *Proceedings of the National Academy of Sciences* 117.30, pp. 17491–17498. ISSN: 0027-8424. DOI: 10.1073/pnas.2003162117.

Heinzelman, Jessica and Waters, Carol (2010). *Crowdsourcing crisis information in disaster-affected Haiti*. US Institute of Peace Washington, DC.

Huynh, Trung Dong, Ebden, Mark, Fischer, Joel, Roberts, Stephen, and Moreau, Luc (2018). "Provenance network analytics". In: *Data Mining and Knowledge Discovery* 32.3, pp. 708–735.

Jansen, Tom, Geleijnse, Gijs, Van Maaren, Marissa, Hendriks, Mathijs P., Ten Teije, Annette, and Moncada-Torres, Arturo (2020). "Machine Learning Explainability in Breast Cancer Survival". In: *30th Medical Informatics Europe Conference, MIE 2020*. IOS Press, pp. 307–311.

Jennings, Nicholas R., Moreau, Luc, Nicholson, David, Ramchurn, Sarvapali D., Roberts, Stephen, Rodden, Tom, and Rogers, Alex (2014). "Human-agent collectives". In: *Communications of the ACM* 57.12, pp. 80–88.

Kraus, Sarit, Azaria, Amos, Fiosina, Jelena, Greve, Maike, Hazon, Noam, Kolbe, Lutz, Lembcke, Tim-Benjamin, Müller, Jörg P, Schleibaum, Sören, and Vollrath, Mark (2020). "AI for explaining decisions in multi-agent environments". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 13534–13538.

Lepri, Bruno, Oliver, Nuria, and Pentland, Alex (2021). "Ethical machines: The human-centric use of artificial intelligence". In: *iScience* 24.3, p. 102249. ISSN: 2589-0042. DOI: https://doi.org/10.1016/j.isci.2021.102249.

Licklider, J. C. R. (1960). "Man-Computer Symbiosis". In: *IRE Transactions on Human Factors in Electronics* HFE-1.1, pp. 4–11. DOI: 10.1109/THFE2.1960.4503259.

Lin, Raz and Kraus, Sarit (2010). "Can automated agents proficiently negotiate with humans?" In: *Communications of the ACM* 53.1, pp. 78–88.

McInerney, James, Stein, Sebastian, Rogers, Alex, and Jennings, Nicholas R. (2013). "Breaking the habit: measuring and predicting departures from routine in individual human mobility". In: *Pervasive and Mobile Computing* 9.6, pp. 808–822.

Moreau, Luc, Freire, Juliana, Futrelle, Joe, McGrath, Robert E., Myers, Jim, and Paulson, Patrick (2008). "The open provenance model: An overview". In: *Proceedings of the International Provenance and Annotation Workshop*. Springer, pp. 323–326.

National Security Commission on AI (2020). *Key Considerations for Responsible Development and Fielding of Artificial Intelligence*. URL: https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-for-Responsible-Development-Fielding-of-AI.pdf.

Neff, Gina and Nagy, Peter (2016). "Automation, Algorithms, and Politics— Talking to Bots: Symbiotic Agency and the Case of Tay". In: *International Journal of Communication* 10.0. ISSN: 1932-8036.

Nisan, Noam, Roughgarden, Tim, Tardos, Eva, and Vazirani, Vijay V. (2007). *Algorithmic Game Theory*. Cambridge University Press. DOI: 10.1017/CBO9780511800481.

Norheim-Hagtun, Ida and Meier, Patrick (2010). "Crowdsourcing for crisis mapping in Haiti". In: *Innovations: Technology, Governance, Globalization* 5.4, pp. 81–89.

Office for Artificial Intelligence (2020). *Guidelines for AI Procurement: a summary of best practices addressing specific challenges of acquiring Artificial Intelligence in the public sector*. URL: https://www.gov.uk/government/publications/guidelines-for-ai-procurement.

Pearl, Judea (2019). "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3, pp. 54–60.

Perez, Caroline Criado (2019). *Invisible women: Exposing data bias in a world designed for men*. Random House.

Pink, Daniel H. (2019). *When: The scientific secrets of perfect timing*. Penguin Press.

Rahwan, Iyad, Cebrian, Manuel, Obradovich, Nick, Bongard, Josh, Bonnefon, Jean-François, Breazeal, Cynthia, Crandall, Jacob W., Christakis, Nicholas A., Couzin, Iain D., Jackson, Matthew O., et al. (2019). "Machine behaviour". In: *Nature* 568.7753, pp. 477–486.

Rahwan, Iyad, Ramchurn, Sarvapali D., Jennings, Nicholas R., McBurney, Peter, Parsons, Simon, and Sonenberg, Liz (2003). "Argumentation-based negotiation". In: *The Knowledge Engineering Review* 18.4, pp. 343–375.

Ramchurn, Sarvapali D., Huynh, Trung Dong, Ikuno, Yuki, Flann, Jack, Wu, Feng, Moreau, Luc, Jennings, Nicholas R., Fischer, Joel E., Jiang, Wenchao, Rodden, Tom, et al. (2015). "HAC-ER: A disaster response system based on human-agent collectives". In: *Proceedings of the 14th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 533–541.

Ramchurn, Sarvapali D., Huynh, Trung Dong, Wu, Feng, Ikuno, Yuki, Flann, Jack, Moreau, Luc, Fischer, Joel E., Jiang, Wenchao, Rodden, Tom, Simpson, Edwin, et al. (2016). "A disaster response system based on human-agent collectives". In: *Journal of Artificial Intelligence Research* 57, pp. 661–708.

Ramchurn, Sarvapali D., Wu, Feng, Jiang, Wenchao, Fischer, Joel E., Reece, Steve, Roberts, Stephen, Rodden, Tom, Greenhalgh, Chris, and Jennings, Nicholas R. (2016). "Human–agent collaboration for disaster response". In: *Autonomous Agents and Multi-Agent Systems* 30.1, pp. 82–111.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos (2018). "Anchors: High-Precision Model-Agnostic Explanations". In: *AAAI-18*, pp. 1527–1535.

Rodden, Tom A., Fischer, Joel E., Pantidi, Nadia, Bachour, Khaled, and Moran, Stuart (2013). "At home with agents: exploring attitudes towards future smart energy infrastructures". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1173–1182.

Salmon, Paul M., Walker, Guy H., and Stanton, Neville A. (2016). "Pilot error versus sociotechnical systems failure: a distributed situation awareness analysis of Air France 447". In: *Theoretical Issues in Ergonomics Science* 17.1, pp. 64–79. DOI: 10.1080/1463922X.2015.1106618. URL: https://doi.org/10.1080/1463922X.2015.1106618.

Samek, Wojciech, Wiegand, Thomas, and Müller, Klaus-Robert (2017). "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *arXiv preprint arXiv:1708.08296*.

Shann, Mike, Alan, Alper, Seuken, Sven, Costanza, Enrico, and Ramchurn, Sarvapali D. (2017). "Save Money or Feel Cozy? A Field Experiment Evaluation of a Smart Thermostat That Learns Heating Preferences". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '17. São Paulo, Brazil: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1008–1016.

Shneiderman, Ben (2020). "Human-centered artificial intelligence: Three fresh ideas". In: *AIS Transactions on Human-Computer Interaction* 12.3, pp. 109–124.

Simon, Herbert A. (1955). "A behavioral model of rational choice". In: *The Quarterly Journal of Economics* 69.1, pp. 99–118.

Simpson, Edwin and Roberts, Stephen (2015). "Bayesian methods for intelligent task assignment in crowdsourcing systems". In: *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*. Springer, pp. 1–32.

Slack, Dylan, Hilgard, Sophie, Jia, Emily, Singh, Sameer, and Lakkaraju, Himabindu (2020). "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, pp. 180–186. ISBN: 9781450371100. DOI: 10.1145/3375627.3375830.

Smith, Virginia, Chiang, Chao-Kai, Sanjabi, Maziar, and Talwalkar, Ameet S. (2017). "Federated Multi-Task Learning". In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4424–4434.

Stein, Sebastian, Ochal, Mateusz, Moisoiu, Ioana-Adriana, Gerding, Enrico H., Ganti, Raghu, He, Ting, and La Porta, Tom (2020). "Strategyproof reinforcement learning for online resource allocation". In: *AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1296–1304.

Tambe, Milind (2011). *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge University Press.

Truong, Ngoc Cuong, Baarslag, Tim, Ramchurn, Sarvapali D., and Tran-Thanh, Long (2016). "Interactive scheduling of appliance usage in the home". In: *25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 869–875.

Tversky, Amos and Kahneman, Daniel (1981). "The framing of decisions and the psychology of choice". In: *Science* 211.4481, pp. 453–458.

Vasconcelos, Wamberto W., Kollingbaum, Martin J., and Norman, Timothy J. (2009). "Normative conflict resolution in multi-agent systems". In: *Autonomous Agents and Multi-agent Systems (JAAMAS)* 19.2, pp. 124–152.

Västberg, Oskar Blom, Karlström, Anders, Jonsson, Daniel, and Sundberg, Marcus (2020). "A Dynamic Discrete Choice Activity-Based Travel Demand Model". In: *Transportation Science* 54.1, pp. 21–41. DOI: 10.1287/trsc.2019.0898.

Venanzi, Matteo, Guiver, John, Kazai, Gabriella, Kohli, Pushmeet, and Shokouhi, Milad (2014). "Community-Based Bayesian Aggregation Models for Crowdsourcing". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. Seoul, Korea: Association for Computing Machinery, pp. 155–164. ISBN: 9781450327442. DOI: 10.1145/2566486.2567989. URL: https://doi.org/10.1145/2566486.2567989.

Verame, Jhim Kiel M., Costanza, Enrico, Fischer, Joel, Crabtree, Andy, Ramchurn, Sarvapali D., Rodden, Tom, and Jennings, Nicholas R. (2018). "Learning from the veg box: designing unpredictability in agency delegation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Verame, Jhim Kiel M., Costanza, Enrico, and Ramchurn, Sarvapali D. (2016). "The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4908–4920.

Weardale, Lord Evans of (2020). *Artificial Intelligence and Public Standards*. Tech. rep. Committee on Standards in Public Life.

Wilson, H. James and Daugherty, Paul R. (2018). "Collaborative intelligence: humans and AI are joining forces". In: *Harvard Business Review* 96.4, pp. 114–123.

Wu, Bichen, Wan, Alvin, Yue, Xiangyu, and Keutzer, Kurt (2018). "SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud". In: *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pp. 1887–1893. DOI: 10.1109/ICRA.2018.8462926.

Yazdanpanah, Vahid, Gerding, Enrico H., Stein, Sebastian, Dastani, Mehdi, Jonker, Catholijn M., and Norman, Timothy (2021). "Responsibility Research for Trustworthy Autonomous Systems". In: *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'21)*, pp. 57–62.

School of Electronics and Computer Science, University of Southampton
E-mail address: sdr1@soton.ac.uk, ss2@ecs.soton.ac.uk

Department of Computing, Imperial College London
E-mail address: nicholas.jennings@imperial.ac.uk