DOI:
10.3233/AIC-220133

Document Version
Peer reviewed version

Multi-Agent Systems Research at King's College London

Elizabeth Black [a], Martim Brandão [a], Oana Cocarascu [a], Bart De Keijzer [a], Yali Du [a], Michael Luck [a,*], Derek Long [a], Albert Meroño-Peñuela [a], Peter McBurney [a], Simon Miles [a], Sanjay Modgil [a], Luc Moreau [a], Maria Polukarov [a], Odinaldo Rodrigues [a] and Carmine Ventre [a]

[a] *Department of Informatics, King's College London, Bush House, 30 Aldwych, London WC2B 4BG, United Kingdom*

E-mail: michael.luck@kcl.ac.uk

Abstract. Current work on multi-agent systems at King's College London is extensive, though largely based in two research groups within the Department of Informatics: the Distributed Artificial Intelligence (DAI) thematic group and the Reasoning & Planning (RAP) thematic group. DAI combines AI expertise with political and economic theories and data, to explore social and technological contexts of interacting intelligent entities. It develops computational models for analysing social, political and economic phenomena to improve the effectiveness and fairness of policies and regulations, and combines intelligent agent systems, software engineering, norms, trust and reputation, agent-based simulation, communication and provenance of data, knowledge engineering, crowd computing and semantic technologies, and algorithmic game theory and computational social choice, to address problems arising in autonomous systems, financial markets, privacy and security, urban living and health. RAP conducts research in symbolic models for reasoning involving argumentation, knowledge representation, planning, and other related areas, including development of logical models of argumentation-based reasoning and decision-making, and their usage for explainable AI and integration of machine and human reasoning, as well as combining planning and argumentation methodologies for strategic argumentation.

Keywords: argumentation, norms, agent-based simulation, strategic interaction, multi-agent reinforcement learning, dialogue protocols

1. Introduction

King's College London has a long history in multi-agent systems, with research dating back to the early 2000s. Current work on multi-agent systems at King's is extensive with many researchers, though largely based in two research groups within the Department of Informatics: the Distributed Artificial Intelligence (DAI) thematic group and the Reasoning & Planning (RAP) thematic group.

DAI combines AI expertise with political and economic theories and data, to explore social and technological contexts of interacting intelligent entities. The group develops computational models for analysing social, political and economic phenomena to improve the effectiveness and fairness of policies and regulations, and combines intelligent agent systems, software engineering, norms, trust and reputation, agent-based simulation, communication and provenance of data, knowledge engineering, crowd computing and semantic technologies, and algorithmic game theory and computational social choice, to address problems arising in autonomous systems, financial markets, privacy and security, urban living and health.

RAP conducts research in symbolic models for reasoning involving argumentation, knowledge representation, planning, and other related areas. Its research priorities are: (i) development of classical, temporal and hybrid planners for constructing efficient plans for complex systems; (ii) development of logical models of argumentation-

*Corresponding author. E-mail: michael.luck@kcl.ac.uk.

based reasoning and decision-making, and their usage for explainable AI and integration of machine and human reasoning; (iii) combining planning and argumentation methodologies for strategic argumentation; (iv) the formalisation of models of ethical reasoning and deliberation; and (v) development of machine learning techniques for topic classification, relationship and argument extraction from unstructured natural language texts.

Beyond these specific groups, there is also a more general focus on autonomous system at King's. Indeed, while there is great interest in deploying autonomy in existing and new applications, the concerns that remain about the safety and trustworthiness of current autonomous systems have led to the establishment of a cross-cutting Trusted Autonomous Systems hub focussing on safe and trusted AI, explainable AI and ethics. It provides a valuable portfolio of skills and expertise on model-based reasoning that can be used to achieve safety of, and trust in, autonomous systems via consideration of causality, explanations (especially for planning and cyber-security), and provenance. In this context, King's is home to the UKRI Centre for Doctoral Training in Safe & Trusted AI, for which multi-agent systems is a key theme, running from 2019 through 2027 as part of the UK Government's investment of £100m initiative to train 1000 new scientists and engineers in AI. King's is also one of three partners in the UKRI Trustworthy Autonomous Systems Hub, which sits at the centre of a £33M Trustworthy Autonomous Systems research programme.

There are also other areas of activity in multi-agent systems at King's. For example, the Department of Engineering's Centre for Robotics Research has the mission to develop solutions to critical challenges faced in society where robot-centric approaches can improve outcomes, including in work on machine learning and on human-robot interaction. Similarly, there has been work undertaken at the King's Institute of Psychiatry, Psychology & Neuroscience on knowledge representation, machine learning, and multi-agent systems applied to biological knowledge discovery, healthcare delivery, and e-health infrastructures, and there is a long tradition of work on multi-agent systems, decision support systems and machine learning applied to clinical trials and biomedical systems in Population Health Sciences. Nevertheless work in Informatics at King's in the groups above provides the critical mass of MAS research, and they remain the focus of this article.

In each key area, the article provides a separate section describing the technical problems, the main approaches & key results, and the open challenges. Given the extent of the variety of work at King's, and for the sake of brevity, we have not attempted to artificially introduce links between the sections, but the natural flow should make clear the relevance and integration to the wider agenda of each. The paper begins with a detailed consideration of argumentation and dialogue, for which King's has one of the most established groupings of researchers, and continues with more foundational examination of agent communications languages and dialogue protocols for machine-to-machine communication. Next the paper covers wider agent interactions, first through work on agent-based models and simulation, then strategic interaction between multiple agents and multi-agent reinforcement learning. Finally, we bring together some related strand of activity to contextualise the work previously outlined.

2. Argumentation and Dialogue

2.1. Technical Problems

A key problem addressed by the group at King's is that of distributed non-monotonic reasoning. The last two decades have witnessed a waning of interest in non-monotonic logics for agent reasoning, due in large part to the unacceptable computational demand exacted when reasoning non-monotonically over realistically complex and rich belief bases, and the increasing dominance and success of machine learning applications that has reinforced scepticism as to whether the symbolic AI paradigm — and more specifically non-monotonic logics — represent the most promising approach to implementation of single agent reasoning.

While it is conceivable that non-monotonic logics may play only a limited role in single agent reasoning (although many convincingly argue that integration of symbolic and machine learning systems will be required for more complex forms of agent reasoning [56]), it is far less contentious to claim that symbolic, and in particular non-monotonic, logics will be required to support joint reasoning among human and artificial agents. After all, more complex reasoning tasks, such as those that involve managing uncertainty and resolving conflicts, often benefit from input and insights elicited from multiple agents engaged in the dialogical exchange of locutions. Hence, non-monotonic logics

can provide normative guidance for rational joint deliberation among human agents, and among human and artificial agents, where the latter will *necessarily* need to communicate symbolically, in ways understandable to their human interlocutors [62]. Indeed, the latter may be of particular importance if AI reasoning and decision making is to be aligned with human values [62, 64][1]. Ensuring that such dialogical exchanges comply with rational principles governing the handling of uncertainty and resolution of conflicts, requires development of formal models of distributed non-monotonic reasoning that can serve to constrain and guide the choice of locutions and their relationships in such dialogues. It is arguably this requirement that is best served by argumentation-based characterisations of non-monotonic logics.

The group at King's has a strong focus on the application of computational argumentation techniques to support distributed non-monotonic reasoning. The problems addressed by the group in this space span the breadth of key challenges and can broadly be classified into three main areas, as follows.

- *Knowledge representation, reasoning, and engineering*, including: development of both structured and abstract argumentation frameworks and argumentative (dialectical) formalisations of non-monotonic logics that comply with rational principles regarding the handling of uncertainty and resolution of conflict (e.g., [22–24, 35, 36, 39, 60, 65, 67, 68, 73, 99]); proof theories and algorithms for reasoning with and about arguments (e.g., [66, 78–80]; specification of argument schemes specialised to particular domains (e.g., [86, 92]); argument mining approaches for identifying arguments and relations between arguments from text (e.g., [16, 18]).
- *Dialogical argumentation*, including: systems for distributed inquiry over beliefs and for distributed deliberation over actions (e.g., [4, 6]); how an agent can act strategically so as to influence the outcome of dialogical argumentation, taking into account what it believes to be true about its interlocutor's private mental state (e.g., [5, 41, 46, 69]).
- *Methods to support human-machine reasoning*, including: techniques for generating argumentation-based interactive recommendations and explanations (e.g., [17]); mechanisms for handling 'enthymemes' (logically incomplete arguments) (e.g., [7, 61, 97]);

2.2. Main Approaches and Key Results

A key knowledge representation challenge is how to represent a belief base to allow the construction of arguments (essentially defeasible proofs) for claims (conclusions), and the identification of relationships between those arguments (such as contradiction-based attack relationships). The *ASPIC+* framework for structured argumentation [68] is a community standard that has been widely studied and developed, and that has established argumentative characterisations of non-monotonic logics (e.g., [99]). However, while *ASPIC+* establishes that these argumentative characterisations yield rational outcomes, it does so under the assumption of logical omniscience; i.e., that agents have unbounded resources. *ASPIC+* also assumes that information about preferences and values that is used to arbitrate among arguments, is fixed and given, and not itself subject to reasoning and disagreement. Moreover, *ASPIC+* assumes an argument ontology that is more suitable for single agent reasoning rather than distributed reasoning. More recently, a novel *dialectical* approach to structured argumentation develops both special cases of *ASPIC+* [22, 23] and the full *ASPIC+* framework [24], so as to: (i) adopt an ontology for arguments that renders it more suitable for distributed reasoning; (ii) establish that rationality is satisifed while making only minimal assumptions on computational resources; and (iii) accommodate arguments constructed in recently proposed depth-bounded logics that satisfy principles of bounded rationality [21].

In order to reason about the justified conclusions of an argumentative belief base, one typically abstracts away from the content of arguments, yielding a graph that represents arguments and their attacks, and applies what are known as 'semantics' [29] to determine the 'winning' arguments. This form of reasoning has a high computational complexity, meaning advanced techniques are needed to overcome implementation issues. Work carried out at King's by Rodrigues and colleagues addresses this challenge in two fronts: the development of alternative numerical semantics to complement or replace the traditional three-valued Dung's semantics, e.g., probabilistic, numerical, and Bayesian semantics as in [34–36] and the investigation of the limits of operations on abstract argumentation

[1] Cooperative Reinforcement Learning [40] – amongst the most promising approaches to ensuring that the values of machines are aligned with human values – points to the need for human and AI agents to engage in dialogical exchange.

frameworks [73]; and the actual design and implementation of algorithms to compute or assist in the computation of solutions to problems in abstract argumentation [77–80].

Away from using abstract argumentation as a technique to characterise non-monotonic reasoning, related work was also done by Rodrigues in terms of logical foundations for knowledge representation and reasoning, more specifically in the areas of *belief revision* [32, 76, 81], *belief contraction* [37, 38], and translation of operations across logical systems [33, 82].

In relation to the practical implementation of agents who reason argumentatively, a key challenge is that of knowledge engineering: where do agents get their arguments from and how do they identify arguments that attack, or support, these? The field of argument mining aims to extract these elements from text and focuses on two main tasks: argument component detection, which aims to identify arguments within text (i.e., claim, premises, and their textual boundaries) and relation prediction of the arguments previously identified. Cocarascu's work focuses on developing machine learning models for argument mining [16, 18] and methods for mining argumentation frameworks from text that can be used to support real-world applications [19].

Also addressing the knowledge engineering challenge, Modgil's work considers the development of argument schemes and critical questions (*AsCq*) specialised for different domains, including organ transplant coordination and medical reasoning more generally (e.g., [86, 92]). Argument schemes are stereotypical patterns of reasoning, which are used as presumptive justification for generating arguments. They are essentially templates that can in principle be instantiated by human and machine interlocutors, and thus potentially provide a bridging level of representation that enables human-computer dialogue. Moreover, each scheme has an associated set of critical questions, which allow one to identify potential attacks on an argument generated by the scheme, where the attacks can then be realised through instantiation of other schemes. In this way, the use of *AsCq* can yield argument graphs of attacking arguments that can be evaluated to determine the outcome of a dialogue.

An important focus is on dialogical argumentation to support distributed reasoning, where agents seek to share relevant arguments in order to reach some joint agreement on what to believe or what to do. In the paradigmatic case of collaborative inquiry over beliefs, it is important to ensure that the outcome is warranted by the joint beliefs of the participants; this property is guaranteed by the inquiry dialogue system developed by Black and Hunter [6], where the participating agents exchange beliefs so as to jointly construct all arguments that may be relevant to the reasoning. However, when deliberating over action, where the best outcome depends on an one's subjective preferences, agents should strategically select arguments to share that are likely to influence the outcome in their favour [4].

Strategic selection of arguments is key whenever an agent aims to influence the result of dialogical argumentation towards a particular outcome. This is a challenging problem, which typically involves consideration of what the strategiser believes about the private mental state of their interlocutor(s) [8]. Black's work with colleagues considers how the problem of determining a strategy that will be effective against one's interlocutor(s) can be mapped to an optimisation problem, so that techniques such as automated planning and evolutionary search can be exploited to identify such a strategy [5, 69].

Arguments presented by humans are normally enthymemes, which are incomplete arguments that omit some of the content of the intended complete argument. To support dialogical argumentation between artificial and human agents, we need to be able to handle enthymemes, and work at King's has have developed a novel set of locutions for managing enthymemes that allow agents to recover from the range of misunderstandings that may arise as a result [97]; previous work of Black (in collaboration with Hunter) considers how an agent can reconstruct the intended argument from a received enthymeme, and can select an enthymeme that will be correctly reconstructed by its recipient, so as to avoid such misunderstandings [7].

Various approaches have been proposed for explainable artificial intelligence (XAI), including argumentation-based explanations, and different types of explanations based on argumentation frameworks can be formed, e.g., conversational or visual. These can then provide the basis for explanations to users by means of interactions. One area in which argumentative explanations have been used by Cocarascu is recommender systems, where argumentation frameworks are extracted from reviews and provide the backbone for dialogical explanations that describe the strengths of movies [17].

2.3. Open Challenges

While *dialectical argumentation* enables formalisation of non-monotonic logics that are rational under resource bounds, it still assumes fixed preferences/values. It remains to further develop the approach to accommodate reasoning about possibly conflicting preferences/values (an important requirement given the above anticipated applications of distributed reasoning; see below). This research goal will be pursued using the same methodology that generalises *ASPIC+* to accommodate reasoning of this kind [60, 67].

In terms of argumentation semantics and computation of semantics of argumentation frameworks, interesting avenues for future work include the automatic extraction of arguments from text expressed in natural language, the representation of these arguments as a formal mathematical structure, and new semantics to enable reasoning over these large quantities of data. The challenges in the development of these new argumentation semantics arise from mechanisms to deal with the uncertainty about the strength of arguments (or their reliability) and how to make sense of an argumentation framework supplied with this extra information. Likewise, as the size of the argumentation frameworks increase, there are technological challenges to the implementation of data structures and algorithms for large scale argumentation graphs and the computation of the enumeration of extensions and the acceptance of arguments under particular semantics.

Much of the work on identifying argument components as well as the pro and con arguments rely on deep learning techniques. One of the main challenges of argument mining is represented by the lack of sufficiently large annotated data that can be used to train machine learning models. To address this, various datasets spanning several domains, have been created [13]. However, there is no agreement or consistency in terms of the annotations among the datasets. Additionally, identifying argument structures in text is a difficult task for humans as well as for machines. Arguments are sometimes determined by the presence of discourse markers, but in most cases, the premise does not follow the claim immediately in the discourse, making it difficult to link them automatically. Furthermore, humans make use of common sense or background knowledge to construct or to identify arguments.

Various approaches to argumentation-based reasoning employ preferences over arguments to determine whether one argument 'successfully attacks' (i.e., *defeats*) another, effectively formalising the use of priorities in non-monotonic logics so as to preferentially arbitrate among conflicting inferences. For example, the well known non-monotonic *Preferred Subtheories* formalism exploits a given total ordering over a belief base so as to define a non-monotonic consequence relation. In argumentative characterisations of Preferred Subtheories inference ([23, 68]), the total ordering is 'lifted' to a preference relation over arguments, so that the claims of justified arguments defined by the belief base, correspond to the Preferred Subtheories inferences defined directly over the belief base. However, one needs to accommodate non-monotonic reasoning *about* priorities/preferences, clearly a more salient requirement when participants in dialogues may differ in respect of their valuation of arguments. An open challenge is therefore to develop mature models of argumentation-based dialogue that accommodate reasoning about preferences (for example, when agents differ as to the relative importance of the values promoted by arguments justifying actions in deliberation dialogues), so as to yield accounts of distributed non-monotonic reasoning that eschew the assumption of exogenously given fixed priorities/preferences. A promising approach to addressing this challenge would be further development of a proposal [63] that accommodates exchange of (possibly conflicting) arguments justifying preferences over other arguments, and where the arguments exchanged are evaluated in *Extended Argumentation Frameworks* [60, 67] that accommodate such arguments via attacks on attacks[2]. The long term aim is to thus develop dialogical, communicative accounts of distributed non-monotonic reasoning that build on the aforementioned extension of *Dialectical Argumentation*, so as to accommodate real-world modes of dialectical exchange, reasoning about preferences/values, and that yield rational outcomes given the pragmatic assumption that real-world agents have bounded resources.

While argumentation has previously been used as a mechanism for providing explanations, more work is needed to determine the right explanation format, which can easily be understood by users. Further work is also required on investigating the incorporation of user feedback into the explanation system.

Finally, further work is required on developing domain specific Schemes and Critical questions (*ScCq*), that will be required to support human-machine dialogue (or indeed to enable computational scaffolding of human-

[2] If X claims that B is preferred to A, then X attacks the attack from A to B.

human dialogue) (see Section 2.2). For example, the anticipated use of such dialogues in supporting alignment of computational decisions with human values [62, 64] can be facilitated by development of *ScCq* that characterise stereotypical patterns of ethical/moral reasoning.

3. Agent Communications Languages and Dialogue Protocols for Machine-to-Machine Communication

3.1. Technical Problems

The problem addressed by McBurney and colleagues is the design of languages and protocols for the automated generation and automated reception of communicative messages between autonomous software entities, or agents. With the growth of the semantic web, the vast majority of messages between computers now are automated requests for digital objects, and automated replies to these requests, using the Hyper-Text Transfer Protocol (HTTP). Considered from the perspective of human argumentation and dialogue theory, these requests are all quite simple: most current machine dialogues would be considered as an *action-requesting* dialogue or, very often, an *information-seeking* dialogue [96]. The research agenda here is to design formal languages and protocols to enable machines to have more sophisticated dialogues than these with one another.

3.2. Main Approaches and Key Results

To achieve this agenda requires formal study of communications and dialogues in terms of their semantics and pragmatics. Semantics refers to the relationship between communications in a formal language and that aspect of reality that concerns the truth of messages, while pragmatics refers to the relationship between communications and other aspects of reality, for example in the way in which messages in a dialogue are used to create or maintain social relationships between the participants. A great deal of effort on computer protocol design has considered the formal syntax of messages and interactions between machines, and the properties of dialogues that arise from these, with some attention having been paid to semantics but usually ignoring pragmatics [57].

In this research endeavour, Shannon's famous theory of communication is of no use, since it explicitly ignores the semantics of messages, and implicitly ignores their pragmatics [88]. Instead, McBurney draws on work in theoretical linguistics (of human communications), the philosophy of argument, and the philosophy of language. Philosophers these last three thousand years, have spent most of their efforts studying propositions, statements that purport to make factual representations about the world.[3] Many interactions, perhaps even most, between humans or between machines are not propositional in nature, however, but involve utterances over actions such as requests, promises, commands, etc. It is the semantics and pragmatics of protocols for dialogue over action that are primarily the focus here.

An example of this is work on exploration of deception in communications: understanding what it is, how it may be engineered so that machines may deceive, and how deception may be recognised and countered by other machines [84]. Applications here are in every form of robot-to-robot communication between machines unknown to each other, where mutual trust can only be assumed by the participants at great risk to themselves. Similarly, deception has a long history of application in military strategy and espionage, as well as in the public disinformation campaigns that have been a feature of international relations since the early days of the Cold War. In addition to work in philosophy, this research has drawn on frameworks for modelling lies and their detection from psychology and ideas from pheology, which has a long history of the study of lying, equivocation and prevarication. The main methods here have involved the modelling of nested theories of mind, and their computational simulation [85].

[3] Exceptions were Thomas Reid [87] and Adolph Reinach [74].

3.3. Open Challenges

Challenges that are still open are: to classify all types of communicative dialogue; to design effective protocols for these different types (particularly those involving actions); to better understand the formal and applied properties of these protocols and their semantics and pragmatics; and to connect these protocols to machine-tractable models of action. People in the blockchain world are fond of saying that smart contracts (automated programs that run on distributed ledgers) can replace natural language legal contracts and thereby enable companies to operate automatically without any human intervention. However, there is currently a very large gap between the subtle and nuanced actions and conversations enabled by commercial natural language contracts and the limited, blunt effects on-chain generated by smart contracts [50]. Research on agent communications protocols over actions will be necessary to bridge this gap.

4. Norms and Behaviour Regulation

4.1. Technical Problems

In systems of interacting autonomous agents, some form of system management or behaviour regulation may be needed, either through constraints imposed by organisational structure and norms (limiting what is possible for agents to do) or through analysis of trust in, and reputation of, potential cooperation partners [49]. When constraints are imposed by organisational structure and norms, trust may be less important, since system regulation achieves compliance. This is the approach adopted in *electronic institutions* [26] in which agents do not have the possibility of violating norms. Yet if agents are less willing to trust others, then the possibility for taking advantage of opportunities in terms of cooperation may be ruled out due to an excessive tendency to caution even if merited by the presence of malicious agents. Balancing these aspects can be crucial in enabling effective systems and societies: Fitoussi and Tennenholtz [31] suggest that norms must be sufficiently restrictive to have the desired effect, but must also be sufficiently flexible so that all objectives are equally feasible.

Since agents are autonomous, compliance with norms is not guaranteed.To encourage compliance and enforce norms, therefore, sanctions may be imposed on a norm violator and agents must consider the possibility of receiving some punishment in the case of violation. For example, in peer-to-peer systems agents share resources with each other, but if there is no cost to accessing resources provided by others, there is also no incentive for agents to contribute their own resources for the benefit of others. More generally, when self-interested autonomous agents exchange information (for example) without central control, non-compliance (due to selfish interests) can compromise the entire system; the nature and volume of interactions can make it impossible to effectively enforce compliance via legal norms. In contrast, social norms are those that emerge through interactions, and are maintained by the individuals that participate within them.

Axelrod, however, showed that norms alone may not lead to the desired outcomes, and that *metanorms* to help enforce compliance of primary norms are required [2]. Yet this also raises questions of how agents should encourage compliance through enforcement and the severity of sanctions and, in open multi-agent systems, how agents should seek to optimise their choice of the most reliable interaction partner among many possible available.

4.2. Main Approaches and Key Results

Axelrod's work has been hugely important and valuable but vastly oversimplifies the problem with too many limiting assumptions. For example, in wireless sensor networks, each agent can only observe the behaviour of a relatively small number of other agents, yet Axelrod assumes that there is a uniform probability of being seen; more generally in real world systems, observability is restricted by network connections rather than some arbitrary probability distribution. In response, work at King's by Mahmoud, Luck and others [53–55] has addressed a number of these constraints, dropping the observability requirement [51] and introducing more realistic topological configurations for considering norm emergence [53], fundamentally changing the mechanisms required to establish cooperation. This allows the development of mechanisms that move beyond a fully connected network, and into such topologies

as lattices, small worlds and scale-free networks (which contain both heavily connected nodes and lightly connected nodes). However, because of the asymmetric nature of scale-free networks with a vast number of connections of hubs and lightly connected outliers, performance is less strong than with other topologies, and a uniform learning rate to modify strategies is ineffective. This has led to examination of adjusting the amount of learning in relation to performance through dynamic policy adaptation [52], bringing about the desired behaviour and norm emergence.

4.3. Open Challenges

Despite the success of these new models building on Axelrod's work, there remain constraints that may prevent some real world application. For example, while observation of interactions may be considered valid in some domains such as social networks (where the network structure determines observability), this could be invalid in other domains (for example, where communication involves some form of encryption, preventing agents from detecting a violation). In addition, resources themselves are often ignored, with the assumption that there is always access to unlimited resources for use in detecting a violation and seeking to enforce norms, but this is clearly not feasible in a world where resources are constrained. More generally, notions of *order and society* from human systems can provide both solutions and inspiration development of techniques in a computational context, yet the interplay of different elements is potentially complex.

5. Engineering for Emergent Behaviour

5.1. Technical Problems

The nature of a multi-agent system means that its population's overall behaviour will emerge from agents' interactions over time and, due to the complexity of the system, this emergence will be unpredictable from the design of individual agents. In order for an engineered agent-based system to be considered reliable, we need to know how it will operate, leading to two key challenges around emergent behaviour. First, if we can anticipate what behaviours could emerge, we could seek to test whether these are likely, but the space of possible future paths the system may take is commonly too large and diverse to feasibly check each one. Second, in a sufficiently rich system, we cannot anticipate every behaviour that may emerge, so we want to instead ensure the system will detect and react to emerging behaviour such that positive behaviour that supports the system's goals is preserved while negative behaviour is eliminated [43]. However, such detection relies on data being exposed by agents about their behaviour and the reasons behind it, and for that data to be reliable, which cannot be assumed in competitive multi-agent systems in which individual agents can gain advantage through obfuscation.

5.2. Main Approaches and Key Results

To question the possible futures of an agent-based system, we can simulate it. Indeed, agent-based simulation is a field of growing importance more generally as it is recognised that purely data-driven modelling is inadequate for analysing complex systems with heterogenous components. It is challenging to determine whether a conclusion drawn from an agent-based model, such as whether a particular behaviour is likely to emerge, is robust or just reflects random chance or the absence of some critical features of the simulated phenomena. Miles and colleagues have developed methods and tools for quantifying whether an observed property can confidently be asserted about the modelled system, drawing on approximate probabilistic model checking and temporal logic [45]. They extended this general approach to allow causal processes within a model to be detected [44].

Addressing the challenge of ensuring reliable data is available from which to detect unanticipated emergent behaviour, the team first built upon their past research in data provenance, providing interoperable standards for capturing causal descriptions of behaviour across distributed systems, and then explored corroboration between agents' accounts as a way of assessing reliability of individual agents' claims [3]. Measuring the reliability of reporting through corroboration does not in itself ensure that agents will behave reliably: there also needs to be a way of incentivising good reporting behaviour. The team extended reputation assessment mechanisms [90] where client

agents rely on the society-constructed reputations of provider agents, and the confidence in reputation scores can be adjusted based on how well corroborated reports are. Where client agents act on behalf of users, the reasoning behind the reputation score can then be explained to them so that they can make an informed choice of provider [71]. In combination, this means that provider agents are incentivised to report behaviour completely and reliably because they will not be chosen to provide future services otherwise.

5.3. Open Challenges

From the above strands of work, there are methodological issues to be solved,and these are the current focus. First, it is important for non-technical domain experts to be able to apply the approach of determining whether an agent-based simulation shows some emergent property or behaviour, because they know the relevant behaviours to check for and how the model might reasonably change. For example, Miles and colleagues work with the emergency department (ED) of a London hospital and have studied how interactions between staff and patients build up to emergent cultural practices that affect the efficacy of the ED over time [9]. Ideally, hospital managers should fully control an agent-based model of the ED so that they can explore alternative ways of working. Here, we require methodologies surrounding our technical approach to make the testing of emergent behaviour accessible and practical.

Within practical scenarios involving emergent behaviour, physical space is an important factor that has not been properly addressed. For example, distance, walls and other obstacles all affect the possibility for interaction and how the system evolves over time (for example, there are only certain places where people can enter or exit the simulated place). The team at King's work with city councils redeveloping public urban spaces for which they wish to ensure that a space's design encourages people to behave in a way that maintains each others' safety, comfort, health, happiness etc. as they interact. A multi-agent system perspective will allow for the psychology of interacting individuals to be properly accounted for in a simulation and analysis, going beyond physical influences on interaction such as trajectory or obstacle avoidance to consider behaviour based on the beliefs or intentions of people in a space.

Finally, when engineering systems to best handle unanticipated emergent behaviour, we need to build upon the reliable detection approaches discussed above to consider how agents can mutually commit to positive emergent behaviour (or an alternative to negative emergent behaviour). This will allow agents not to have to spend resources preparing for multiple contingencies but instead be able to expect and build upon emergent behaviour that has been found to be beneficial for the system as a whole.

6. Strategic Interaction Between Multiple Agents

6.1. Technical Problems

Ventre, Polukarov and De Keijzer study issues around the strategic interaction between multiple agents in a system. Problems of interest include collective decision making, incentive-compatible mechanisms and stability analysis, which are tackled from complementary perspectives on agents' rationality; agents are either assumed to be perfectly rational or to have certain cognitive biases. In the former case, there is a need to develop incentive models to explain, predict or drive an agent's behaviour towards outcomes that are desirable from the system's perspective. In the latter case, there is a focus on providing theoretical guarantees about what systemic outcomes are compatible within the given cognitive limitations. In addition to this, analytical research, work (together with industry partners) has been conducting empirical studies for the strategic response of trading agents to different market microstructure design in highly dynamic financial markets. Other practical applications include, among others, tactical voting and participatory budgeting domains.

6.2. Main Approaches and Key Results

The question of how agents — human or artificial — make collective decisions, is central to both computer and social sciences. In this context, the notion of strategic voting has been highlighted in research on (computational) social choice as crucial to understanding the relationship between preferences of decision-makers and the outcome of elections. To this end, Polukarov and colleagues proposed a paradigm of iterative voting (see, for example, [58]), which focuses on driving strategic voting behaviour towards a mutually agreed joint decision (instead of preventing one). Motivated by web services such as Doodle or Survey Monkey, an iterative voting process starts from some initial (most commonly, the truthful) voting configuration, and lets the agents subsequently make myopic improvements to the outcome by changing their current vote. In so doing, the model embraces the inevitable manipulability of voting mechanisms (i.e., Gibbard-Satterthwaite impossibility) and views the agents' ability to vote strategically as a collective opportunity to reach stable, mutually agreed decisions. Iterative voting has received significant attention in the recent AI and multi-agent systems literature due, in part, to its potential to provide good predictions for the outcome of the election (and hence, the state of the system). For instance, it has been shown that an iterative process can eliminate low quality Nash equilibria that may arise in certain elections and can successfully model the electorate response to poll data even when voters have limited information and restricted communication abilities. Moreover, (algorithmic) game-theoretic analysis is employed to introduce a suitable equilibrium refinement notion [72] that is consistent with the behaviour recorded in real-life data for, e.g., Doodle polls where decisions are not only guided by the intrinsic preferences over the alternatives. This idea of rationalising the real-life agents' behaviour is extended by proposing an alternative solution concept of cooperative equilibrium [14] to capture the phenomena observed in the context of social dilemmas.

Another fundamental aspect of multi-agent decisions relates to resource (or task) allocation domains where agents hold private information about their values for possible allocations. In this context, work by De Keijzer, Ventre and colleagues has built the foundations of (algorithmic) mechanism design for imperfectly rational agents. A host of results have been proven for a particular cognitive limitation to do with the ability of reasoning contingently, see, for example, [1, 25, 30]. In addition, agent-based models have been used, enriched by an empirical game-theoretic analysis to study macro phenomena from the micro-interactions guided by the incentives of the traders in markets. Questions examined include novel market designs improving market behaviour and reducing pernicious market manipulation strategies and herding behaviour and its effects on flash crashes. Novel algorithms have also been devised for solving Büchi games and Parity games, which have applications in model checking, verification, and safety-critical systems. In addition, a collection of two-sided market mechanisms has been designed and analysed, and proven to have high computational efficiency and social welfare guarantees, suitable for various general classes of market settings [20]. The task of finding clearing solutions in financial networks has also been investigated [83], and network structures identified under which irrationality and high computational complexity prevent the computation of exact solutions (alongside algorithms which can solve this task efficiently in some useful special cases). Lastly, there has been an effort to study a dynamic facility reallocation problem, characterising the best way of moving a facility along a linear space so as to optimally trade-off movement cost and proximity to a set of users of the facility who each move over time.

6.3. Open Challenges

The computational social choice literature offers a systematic, in-depth analysis of strategic voting behaviour in the context of single-winner elections — preference aggregation procedures that output a single winning alternative; recently, the investigation of these questions has started in the context of selecting committees (also known as multi-winner elections) which captures a wide range of applications, such as electing political leaders, determining outcomes of talent competitions or hiring procedures, identifying the best items to recommend to a user of online media based on the reported experiences of other users, choosing papers to be presented at a conference, deciding on a set of measures to achieve a particular target (such as reducing carbon emissions or controlling viral transmission) or splitting a limited budget among competing projects. The latter scenario is typical in the context of participatory budgeting, in which city residents decide directly on the distribution of public funds. However, the literature on strategic behaviour in multi-winner elections is scarce and is almost non-existent in the context of participatory budgeting. To this end, and in collaboration with Westminster City Council, we aim to:

- develop a suite of multi-winner (iterative) voting procedures that admit efficient algorithms on realistic inputs and/or convergent dynamics of strategic moves;
- identify a set of guiding principles that can be used to choose an appropriate procedure from this suite for a specific decision-making scenario; and
- identify the ways for these procedures to be used to increase public engagement and/or guide decision-makers towards a desirable decision in a given scenario.

For this agenda to succeed, it is important to accurately account for (or, in other words, rationalise) the behaviours observed in practice. This issue is also critical for building correct incentive models in the context of mechanism design where our main research challenges include: a deeper understanding of mechanism design for imperfect rationality; and, designing and deploying AI agents in financial markets. Within the finance domain, we aim to solve problems concerning: systemic risk in financial networks; and, the design of efficient and transparent platforms/mechanisms for bilateral and multilateral markets. Finally, problems of interest that cut across the research directions above include: improving agent learning in presence of noisy labels; and analysis of voting protocols and estimation of their outcome distribution using Markov chain Monte Carlo techniques.

7. Multi-Agent Reinforcement Learning

7.1. Technical Problems

Multi-agent systems are increasingly ubiquitous, with systems such as traffic light control and autonomous driving becoming ever more prevalent. However, multi-agent reinforcement learning (MARL) is still confined to a small subset of multi-agent systems with limited complexity. Du's research agenda for MARL considers the following challenges.

- How to flexibly control an arbitrary number of agents, such as coordinating a varying number of vehicles at an intersection?
- How to incentivise agents to contribute rather than free-riding when only team reward rather than individual reward is available?
- Beyond competitive and team games, general scenarios in which agents are selfish or self-interested exist more widely, but the literature concerning them is very sparse.
- Evaluation is essential in driving progress in machine learning, but little attention has been paid to evaluations of players compared to developing algorithms. Challenging scenarios here include evaluating human players, intransitive skills, and data efficiency.

7.2. Main Approaches and Key Results

To address these problems, Du and colleagues explore several strands of activity. First, they proposed a novel architecture that learns a spatial joint representation of all the agents and outputs grid-wise actions [42]. Each agent is controlled independently by taking the action from the grid it occupies. The proposed method can be conveniently integrated with general reinforcement learning algorithms, such as PPO and Q-learning, and its effectiveness was demonstrated in extensive challenging multi-agent tasks in StarCraft II. Second, they proposed a bi-level framework with which each agent learns an intrinsic reward function that diversely stimulates the agents at each time step [27]. Empirical results on StarCraft II demonstrate the effectiveness of LIIR, and LIIR can assign each individual agent an insightful intrinsic reward per time step. In addition, Du and colleagues have developed a MARL solver that computes the Nash equilibrium (NE) within a new subclass of stochastic games [59]. Theoretically, the learning method enables independent agents to learn Nash equilibrium strategies in polynomial time, and the framework outperforms the state-of-the-art MARL baselines in tackling tasks such as large selfish routing. Finally, this line of research focuses on reducing the number of pairwise comparisons in recovering a satisfying ranking for players in two-player meta-games, by exploring the fact that agents with similar skills may achieve similar payoffs against others, and by active sampling [28, 98].

7.3. Open Challenges

There are two main open challenges here. First, while some new evaluation methods have been developed, new problems nevertheless keep emerging, such as off-policy evaluation and cooperative evaluation. Second, reinforcement learning is very close to the human cognition process, but questions such as measurement of agents' consciousness and intelligence are neither answered nor properly asked.

8. Other Areas of Multi-Agent Systems Research

The core areas of multi-agent systems research at King's are described above, but there are are many other relevant areas of focus, not least in relation to the data on which agent systems operate, and the increasing need to be able to track and explain their behaviour.

Indeed, while operating with various degrees of autonomy, multi-agent systems make decisions that may affect humans. Such systems may therefore be subject to laws and regulations, such as the potential requirement of explainability in the GDPR [47, 91, 93]. Independently of regulatory needs, there is an increasing desire to be transparent about decisions that affect people, and society demands accountability for outcomes resulting from (semi-) automated processing. The problems of explainability, transparency, and accountability are challenging to address in multi-agent systems because a single agent, by definition, does not have an overarching view over the decentralised decision-making process, and thus is unable to provide a comprehensive explanation about an actual the distributed process that led to a given decision.

Knowledge graphs (KGs) are knowledge bases that use a graph-based data model to capture knowledge in application scenarios that involve integrating, managing and extracting value from diverse sources of data at large scale [70]. Many of them are engineered through the lenses of disciplines that have been traditionally central in AI, such as knowledge representation [94] and semantic networks [89]. The key issue around KGs, therefore, is to represent the knowledge of the world through connected symbols with well-defined meaning in such a way that when humans and machines consume them, they can derive consequent facts in a predictable and sound manner.

Many of the challenges of creating these KGs come from the fact that this 'knowledge of the world' has an inherently unfathomable scale and complexity. To overcome it, traditional knowledge engineering has focused on creating KGs of a reasonably small size to keep these issues at bay; however, with the proliferation of industrial, large-scale KGs (see, for example, Google[4]) new techniques addressing these issues, especially due to the needs of Web users, have become necessary. An effective approach to deal with them, imitating the success of projects like Wikipedia, has been the exploitation of collaborative and peer-production systems and communities. For example, Wikidata [95] is built and maintained in this way, and is today the largest open-source knowledge graph with 100 million entities curated by a community of 24,000 active editors. The consistency of KGs like Wikidata, the logical formalisms needed for enabling reasoning at scale on them, their querying, and supporting multilingual and multimodal (for example, from images, sound, etc.) knowledge are still unresolved challenges.

Being these large scale, knowledge-based representations, KGs can be a powerful tool for multi-agent systems (MAS). For instance, they can provide models of the world that MAS can build upon as either partial (constrained) or complete models of the domain they are deployed at. Moreover, these models can be parametrised with different levels of logical complexity and semantics, enabling MAS to infer new facts from the environment and their actions, planned or executed, through e.g. description logics reasoning and federated querying. With new methodologies leveraging sub-symbolic reasoning in e.g. neural networks, the knowledge in KGs can also be represented as tensors in Euclidean space [75], offering an opportunity for hybrid KG-MAS embeddings for various downstream tasks.

Against this background, provenance in multi-agent systems offers a decentralised mechanism to describe the flow of data leading to a decision, the transformations applied to such data, and the agents responsible for transforming and communicating data. In multi-agent systems, the problem of explaining a decision, or providing a transparent account of an outcome, can be seen as a cooperative process between the various agents involved in the system: agents need to cooperate in order to deliver suitable explanations to the various stakeholders. Research in this area

[4] Amit Singhal, "Introducing the knowledge graph: things, not strings," Official google blog 5 (2012): 16

involves: the design of taxonomies for explanations [12]; the use of provenance as the knowledge representation to describe the data flowing within agents, and exchanged by agents; and cooperation protocols allowing the delivery of explanations to the targeted audience, while ensuring that confidential information is not revealed by agents.

In this context of explanation, Brandão and collaborators have been investigating multi-agent path finding (MAPF) methods [12]. They have been working together with developers and industry users of these methods to understand what kinds of things may need to be explained, what useful explanations would look like, and what requirements and design considerations are needed to develop explanation generation algorithms. Brandão has shown that inverse optimisation is a useful tool for explanation [11], since it allows changes to be found to the original planning problem that would lead a planner to provide the output that was expected by a user. For example, "agent i does not traverse location X because there is an obstacle at location Y (if there was no obstacle at Y then agent i would have traversed X)". How to efficiently do inverse optimisation for explanation in large MAPF problems is still an open challenge. So far, Brandao has shown that incremental inverse optimisation methods can provide drastic speed-ups [10, 11] compared to traditional formulations, but at the cost of incompleteness in particular kinds of explanation problems [10]. Other open challenges in explainable MAS/MAPF include, for example: (i) how to generate explanations for real-world multi-agent planning methods, such as sub-optimal, incomplete, lifelong and anytime planners used in warehouse automation and computer games; (ii) how to identify core events responsible for long-term behaviour; and (iii) how to automatically generate abstractions for explanation at various levels of detail and thus suitable for various types of user needs.

As this article reveals, at King's College London the breadth of research into multi-agent systems is extensive. The paper has focussed on providing an outline of key areas of activity in relation to core elements of the field, while also pointing to important and relevant related areas of focus. From argumentation and dialogue through protocols for machine-machine communication and strategic interaction, from norm-based systems through emergent behaviour in agent-based simulations, and with the addition of knowledge, provenance and explanation, the paper presents a very significant grouping of researchers. It should be noted, however, that there is even more than represented here of interest to the multi-agent systems community. For example, among other areas, work on planning by Long and colleagues addresses (i) centralised planning for multiple agent execution and how to sustain successful execution in the face of possible plan friction or failure; (ii) signposting *commitments* — responsibilities that are laid on each executive in their interactions with the activities of others — including the timing and the precise conditions required in order for the dependent agents to be able to continue with their own execution. Similarly, the real-world application of multi-agent systems in various domains (e.g., [15, 48]) is also a key characteristic of wider work at King's.

References

[1] T. Archbold, B. de Keijzer, and C. Ventre. Non-obvious manipulability for single-parameter agents and bilateral trade. *CoRR*, abs/2202.06660, 2022.

[2] R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.

[3] L. Barakat, P. Taylor, N. Griffiths, and S. Miles. A reputation-based framework for honest provenance reporting. *ACM Transactions on Internet Technology*, 2022. In press.

[4] E. Black and K. Atkinson. Choosing persuasive arguments for action. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pages 849–856, 2011.

[5] E. Black, A. J. Coles, and C. Hampson. Planning for persuasion. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 933–942, 2017.

[6] E. Black and A. Hunter. An inquiry dialogue system. *E. Agents and Multi-Agent Systems*, 19(2):173–209, 2009.

[7] E. Black and A. Hunter. A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55 – 78, 2012.

[8] E. Black, N. Maudet, and S. Parsons. Argument-based dialogue. In *Handbook of Formal Argumentation, Volume 2*, pages 511–576. College Publications, 2021.

[9] O. Boiko, M. Edwards, S. Zschaler, S. Miles, and Rafferty A.-M. Interprofessional barriers in patient flow management: an interview study of the views of emergency department staff involved in patient admissions. *Journal of Interprofessional Care*, 35(3):334–342, 2020.

[10] M. Brandao. 'why not this mapf plan instead?' contrastive map-based explanations for optimal mapf. In *ICAPS 2021 Workshop on Explainable AI Planning (XAIP)*, June 2022.

[11] M. Brandao, A. Coles, and D. Magazzeni. Explaining path plan optimality: Fast explanation methods for navigation meshes using full and incremental inverse optimization. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pages 56–64, Aug 2021.

[12] M. Brandao, M. Mansouri, A. Mohammed, P. Luff, and A. Coles. Explainability in multi-agent path/motion planning: User-study-driven taxonomy and requirements. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2022.

[13] E. Cabrio and S. Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 5427–5433, 2018.

[14] V. Capraro, M. Venanzi, M. Polukarov, and N. R. Jennings. Cooperative equilibria in iterated social dilemmas. In B. Vöcking, editor, *Algorithmic Game Theory - 6th International Symposium, SAGT 2013, Aachen, Germany, October 21-23, 2013. Proceedings*, volume 8146 of *Lecture Notes in Computer Science*, pages 146–158. Springer, 2013.

[15] M. Chapman, P. Balatsoukas, M. Ashworth, V. Curcin, N. Kökciyan, K. Essers, I. Sassoon, S. Modgil, S. Parsons, and E. I. Sklar. Computational argumentation-based clinical decision support. In E. Elkind, M. Veloso, N. Agmon, and M. E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 2345–2347. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[16] O. Cocarascu, E. Cabrio, S. Villata, and F. Toni. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA*, pages 45–52, 2020.

[17] O. Cocarascu, A. Rago, and F. Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems AAMAS*, pages 1261–1269, 2019.

[18] O. Cocarascu and F. Toni. Identifying attack and support argumentative relations using deep learning. In *Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 1374–1379, 2017.

[19] O. Cocarascu and F. Toni. Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, 44(4), 2018.

[20] R. Colini-Baldeschi, P. W. Goldberg, B. de Keijzer, S. Leonardi, T. Roughgarden, and S. Turchetta. Approximately efficient two-sided combinatorial auctions. *ACM Trans. Economics and Comput.*, 8(1):4:1–4:29, 2020.

[21] M. D'Agostino, D.M. Gabbay, and S. Modgil. Normality, non-contamination and logical depth in classical natural deduction. *Studia Logica*, 108(2):291–357, 2017.

[22] M. D'Agostino and S. Modgil. Classical logic, argument and dialectic. *Artificial Intelligence*, 262:15–51, 2018.

[23] M. D'Agostino and S. Modgil. A study of argumentative characterisations of preferred subtheories. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1788–1794, 2018.

[24] M. D'Agostino and S. Modgil. A fully rational account of structured argumentation under resource bounds. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1841–1847. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[25] B. de Keijzer, M. Kyropoulou, and C. Ventre. Obviously strategyproof single-minded combinatorial auctions. In A. Czumaj, A. Dawar, and E. Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPIcs*, pages 71:1–71:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[26] M. d'Inverno, M. Luck, P. Noriega, J. A. Rodríguez-Aguilar, and C. Sierra. Communicating open systems. *Artificial Intelligence*, 186:38–94, 2012.

[27] Y. Du, L. Han, M. Fang, J. Liu, T. Dai, and D. Tao. LIIR: learning individual intrinsic reward in multi-agent reinforcement learning. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4405–4416, 2019.

[28] Y. Du, X. Yan, X. Chen, J. Wang, and H. Zhang. Estimating α-rank from A few entries with low rank matrix completion. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2870–2879. PMLR, 2021.

[29] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[30] D. Ferraioli, P. Penna, and C. Ventre. Two-way greedy: Algorithms for imperfect rationality. In M. Feldman, H. Fu, and I. Talgam-Cohen, editors, *Web and Internet Economics - 17th International Conference, WINE 2021, Potsdam, Germany, December 14-17, 2021, Proceedings*, volume 13112 of *Lecture Notes in Computer Science*, pages 3–21. Springer, 2021.

[31] D. Fitoussi and M. Tennenholtz. Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence*, 119(1–2):61–101, 2000.

[32] D Gabbay, O Rodrigues, and A Russo. *Revision, Acceptability and Context: Theoretic and Algorithmic Aspects*. Springer Verlag, 2010.

[33] D. M. Gabbay, G. Pigozzi, and O. Rodrigues. Belief revision, belief merging and voting. In *Proceedings of the Seventh Conference on Logic and the Foundations of Games and Decision Theory (LOFT06)*, pages 71–78. University of Liverpool, 2006.

[34] D. M. Gabbay and O. Rodrigues. Equilibrium states in numerical argumentation networks. *Logica Universalis*, pages 1–63, 2015.

[35] D. M. Gabbay and O. Rodrigues. Probabilistic argumentation: An equational approach. *Logica Universalis*, 9(3):345–382, 2015.

[36] D. M. Gabbay and O. Rodrigues. Introducing bayesian argumentation networks. *The IfColog Journal of Logics and their Applications*, 3(2):241–278, 2016.

[37] D. M. Gabbay, O. Rodrigues, and J. Woods. Belief contraction, anti-formulae and resource overdraft: Part I - Deletion in resource bounded logics. *Logic Journal of the IGPL*, 10(6):601–652, November 2002.

[38] D. M. Gabbay, O. Rodrigues, and J. Woods. Deletion in resource unbounded logics - Belief contraction, anti-formulae and resource overdraft: Part II. In S. Rahman, J. Symons, D. M. Gabbay, and J. P. van Bendegem, editors, *Logic, Epistemology and the Unity of Science*, volume 1, chapter 16, pages 291–326. Kluwer Academic Publishers, 2004.

[39] D. Grossi and S. Modgil. On the graded acceptability of arguments in abstract and instantiated argumentation. *Artificial Intelligence*, 275(2):138–173, 2019.

[40] D. Hadfield-Menell, A. Dragan, P.Abbeel, and S. Russell. Cooperative inverse reinforcement learning. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3916?3924, 2016.

[41] C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney. Opponent modelling in persuasion dialogues. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 164–170, 2013.

[42] L. Han, P. Sun, Y. Du, J. Xiong, Q. Wang, X. Sun, H. Liu, and T. Zhang. Grid-wise control for multi-agent reinforcement learning in video game AI. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2576–2585. PMLR, 2019.

[43] C. Haynes, M. Luck, P. McBurney, S. Mahmoud, T. Vítek, and S. Miles. Engineering the emergence of norms: a review. *The Knowledge Engineering Review*, 32, 2017.

[44] B. Herd and S. Miles. Detecting causal relationships in simulation models using intervention-based counterfactual analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–25, 2019.

[45] B. Herd, S. Miles, P. McBurney, and M. Luck. Quantitative analysis of multi-agent systems through statistical verification of simulation traces. *International Journal of Agent-Oriented Software Engineering*, 6(2):156–186, 2018.

[46] M.A. Hosseini, S. Modgil, and O. Rodrigues. Assigning likelihoods to interlocutors, beliefs and arguments. In *Proceedings of the 6th International Conference on Computational Models of Argument. (COMMA 2016)*, pages 339–350, 2016.

[47] T. D. Huynh, N. Tsakalakis, A. Helal, S. Stalla-Bourdillon, and L. Moreau. Addressing regulatory requirements on explanations for automated decisions with provenance: A case study. *Digital Government: Research and Practice*, 2(2), January 2021.

[48] Z. M. Ibrahim, L. Fernández de la Cruz, A. K. Stringaris, R. Goodman, M. Luck, and R. J. B. Dobson. A multi-agent platform for automating the collection of patient-provided clinical feedback. In G. Weiss, P. Yolum, R. H. Bordini, and E. Elkind, editors, *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, pages 831–839. ACM, 2015.

[49] M. Luck. Flexible behaviour regulation in agent based systems. In S. A. Dobson, J. Strassner, M. Parashar, and O. Shehory, editors, *Proceedings of the 6th International Conference on Autonomic Computing, ICAC 2009, June 15-19, 2009, Barcelona, Spain*, pages 147–148. ACM, 2009.

[50] D. Magazzeni, P. McBurney, and W. Nash. Validation and verification of smart contracts: a research agenda. *IEEE Computer Journal*, 50:50–57, 2017.

[51] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck. Overcoming omniscience for norm emergence in axelrod's metanorm model. In S. Cranefield, M. B. van Riemsdijk, J. Vázquez-Salceda, and P. Noriega, editors, *Coordination, Organizations, Institutions, and Norms in Agent System VII, COIN 2011 International Workshops, COIN@AAMAS 2011, Taipei, Taiwan, May 3, 2011, COIN@WI-IAT 2011, Lyon, France, August 22, 2011, Revised Selected Papers*, Lecture Notes in Computer Science, pages 186–202. Springer, 2011.

[52] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck. Norm emergence through dynamic policy adaptation in scale free networks. In H. Aldewereld and J. Simão Sichman, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems VIII - 14th International Workshop, COIN 2012, Held Co-located with AAMAS 2012, Valencia, Spain, June 5, 2012, Revised Selected Papers*, volume 7756 of *Lecture Notes in Computer Science*, pages 123–140. Springer, 2012.

[53] S. Mahmoud, N. Griffiths, J. Keppens, and M. Luck. Establishing norms with metanorms over interaction topologies. *Autonomous Agents Multi-Agent Systems*, 31(6):1344–1376, 2017.

[54] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, T. J. M. Bench-Capon, and M. Luck. Establishing norms with metanorms in distributed computational systems. *Artificial Intelligence and Law*, 23(4):367–407, 2015.

[55] S. Mahmoud, S. Miles, and M. Luck. Cooperation emergence under resource-constrained peer punishment. In C. M. Jonker, S. Marsella, J. Thangarajah, and K. Tuyls, editors, *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pages 900–908. ACM, 2016.

[56] G. F. Marcus. Deep learning: A critical appraisal. *ArXiv*, abs/1801.00631, 2018.

[57] P. McBurney and S. Parsons. Dialogue games for agent argumentation. In I. Rahwan and G. Simari, editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer, Berlin, Germany, 2009.

[58] R. Meir, M. Polukarov, J. S. Rosenschein, and N. R. Jennings. Iterative voting and acyclic games. *Artif. Intell.*, 252:100–122, 2017.

[59] D. H. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, and J. Wang. Learning in nonzero-sum stochastic games with potentials. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7688–7699. PMLR, 2021.

[60] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.

[61] S. Modgil. Revisiting abstract argumentation frameworks. In *Theory and Applications of Formal Argumentation (TAFA'14)*, pages 1–15, 2014.

[62] S. Modgil. Dialogical scaffolding for human and artificial agent reasoning. In *Proceedings of the 5th International Workshop on AI and Cognition*, pages 58–71, 2017.

[63] S. Modgil. Towards a general framework for dialogues that accommodate reasoning about preferences. In *Theory and Applications of Formal Argumentation*, pages 175–191, 2017.

[64] S. Modgil. Many kinds of minds are better than one : Value alignment through dialogue. In *Workshop on Argumentation and Philosophy (co-located with COMMA'18)*, 2018.

[65] S. Modgil and T.J. M. Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 21(6):959 – 1003, 2010.

[66] S. Modgil and M. Caminada. Chapter 6 : Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G. Simari, editors, *Argumentation in AI*, pages 105–129. Springer, 2009.

[67] S. Modgil and H. Prakken. Reasoning about preferences in structured extended argumentation frameworks. In *Proceedings of Computational Models of Argument: COMMA 2010*, pages 347–358, 2010.

[68] S. Modgil and H. Prakken. A general account of argumentation and preferences. *Artificial Intelligence*, 195(0):361 – 397, 2013.

[69] J. Murphy, A. Burdusel, M. Luck, S. Zschaler, and E. Black. Deriving persuasion strategies using search-based model engineering. In *Proceedings of the 7th International Conference on Computational Models of Argument*, pages 221–232, 2018.

[70] N. Fridman Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM*, 62(8):36–43, 2019.

[71] I. Nunes, P. Taylor, L. Barakat, N. Griffiths, and S. Miles. Explaining reputation assessments. *International Journal of Human-Computer Studies*, 123:1–17, 2019.

[72] S. Obraztsova, M. Polukarov, Z. Rabinovich, and E. Elkind. Doodle poll games. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, page 876–884, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[73] O. Rodrigues R. Baumann, D. M. Gabbay. Forgetting an argument. In *Proceedings of the 34th Conference on Artificial Intelligence*, 2020.

[74] A. Reinach. Die apriorischen Grundlagen des bürgerlichen Rechtes. *Jahrbuch für Philosophie und phänomenologische Forschung*, 1:685–847, 1913.

[75] Petar Ristoski and Heiko Paulheim. Rdf2vec: RDF graph embeddings for data mining. In Paul Groth, Elena Simperl, Alasdair J. G. Gray, Marta Sabou, Markus Krötzsch, Freddy Lécué, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, volume 9981 of *Lecture Notes in Computer Science*, pages 498–514, 2016.

[76] O. Rodrigues. *Iterated Revision and Automatic Similarity Generation*, volume 2, pages 591–613. College Publications, 2005.

[77] O. Rodrigues. EqArgSolver – System Description. In *Proceedings of the 4th Intl. Conference on Theory and Applications of Formal Argumentation*, TAFA'17, pages 150–158, 2018.

[78] O. Rodrigues. A forward propagation algorithm for the computation of the semantics of argumentation frameworks. In *Theory and Applications of Formal Argumentation*, TAFA'17, pages 120–136. Springer International Publishing, 2018.

[79] O. Rodrigues. An investigation into reduction and direct approaches to the computation of argumentation semantics. In J.-Y. Beziau, A. T. Martins, F. Ferreira, and M. Pequeno, editors, *Logic, Intelligence and Artifices: Tributes to Tarcision H. C. Pequeno*, pages 97–120. College Publications, 2018.

[80] O. Rodrigues. Representing and comparing large sets of extensions of abstract argumentation frameworks. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC'19)*, 2019.

[81] O. Rodrigues, D. Gabbay, and A. Russo. Belief revision. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic: Volume 16*, pages 1–114. Springer Netherlands, Dordrecht, 2011.

[82] O. Rodrigues, D. M. Gabbay, and A. Russo. Belief revision in non-classical logics. *Review of Symbolic Logic*, 1:267–304, 2008.

[83] C. Ventre S. D. Ioannidis, B. de Keijzer. Strong approximations and irrationality in financial networks with derivatives. In *ICALP 2022: 49th International Colloquium on Automata, Languages, and Programming*, 2022, in press.

[84] S. Sarkadi. *Deception*. Ph.d., Department of Informatics, King's College London, London, UK, 2020.

[85] S. Sarkadi, A. R. Panisson, R. H. Bordini, P. McBurney, S. D. Parsons, and M. D. Chapman. Modelling deception using Theory of Mind in multi-agent systems. *AI Communications*, 32:287–302, 2019.

[86] I. Sassoon, N. Kökciyan, S. Modgil, and S. Parsons. Argumentation schemes for clinical decision support. *Argument and Computation*, 12(3):329–355, 2021.

[87] K. Schuhmann and B. Smith. Elements of Speech Act Theory in the work of Thomas Reid. *History of Philosophy Quarterly*, 7:47–66, 1990.

[88] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Chicago, IL, USA, 1963.

[89] John F. Sowa. Semantic networks, 1987.

[90] P. Taylor, L. Barakat, S. Miles, and N. Griffiths. Reputation: A review and unifying abstraction. *Knowledge Engineering Review*, 33, 2018.

[91] The UK Information Commissioner's Office. Explaining decisions made with AI. Technical report, 2020.

[92] P. Tolchinsky, S. Modgil, K. Atkinson, P. McBurney, and U. Cortes. Deliberation dialogues for reasoning about safety critical actions. *Journal of Autonomous Agents and Multi-Agent Systems*, 25:209–259, 2012.

[93] N. Tsakalakis, S. Stalla-Bourdillon, L. Carmichael, T. D. Huynh, L. Moreau, and A. Helal. The dual function of explanations: Why it is useful to compute explanations. *Computer Law and Security Review*, 41, March 2021.

[94] Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors. *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*. Elsevier, Amsterdam, 2008.

[95] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[96] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Series in Logic and Language. State University of New York Press, Albany, NY, USA, 1995.

[97] A. Xydis, C. Hampson, S. Modgil, and E. Black. Enthymemes in dialogue. In *Proceedings of the 8th International Conference on Computational Models of Argument*, pages 395–402, 2020.

[98] X. Yan, Y. Du, B. Ru, J. Wang, H. Zhang, and X. Chen. Learning to identify top elo ratings: A dueling bandits approach. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 6375–6383. AAAI Press, 2022.

[99] A.P. Young, S. Modgil, and O. Rodrigues. Prioritised default logic as rational argumentation. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'16)*, pages 626–634, 2016.