Investigating Human Considerations in State-of-the-Art Reinforcement Learning Models

Adrian Azzarelli, aab1g18@soton.ac.uk

In human-cerntric reinforcement learning (HCRL), applied technologies acts only as a substrate to some interdisciplinary dilemma, yet research often neglects the non-technical aspects of either a problem or its solutions. This is evidenced by novel research into responsible HCRL design, which outlines flaws in current technologies surrounding the lack of *explainability, intepretability* and *transparency* criteria. By analysing both the criticisms and technology we delineate a set of principles which instruct responsible design. We further evaluate a varied sample of recent HCRL research against the criteria to provide insight into characteristics which drive responsible/not so responsible system design. Consequently, we provide educated suggestions towards improving particular aspects of research which can be applied generally.

Human-centric reinforcement learning | Responsible design | Meta-learning

1 Introduction

Reinforcement Learning (RL) provides structure to solving generalised, often seemingly abstract, sets of sequential problems. Consequently, one can find many examples of video games, board games and even realistic situations where RL has been adapted to generalise under a considerably large number of constraints. For example, the game *Go* was famously solved using RL, Silver et al. (2016), despite the arduity associated with the game's "enormous search space".

Solving such problems has been an on-going feat; popularised by the defeat of *Chess* in 1996[1]. Ever since, the ambition to train robots on what is seemingly the abstract notion of playing a "game" has increased. Considering this we contemplate whether such solutions provide us with evidence of functional understanding (as initially intended, Haykin, Network (2004)) or if what was previously just "abstract notion" has been broken down into some generalised representation of the real problem. Simply put, we question whether true optimality can be achieved, as opposed to the accepted pareto optimal solution.

It may be easy to accept a pareto optimal solution, particularly when a logical RL solutions are available. However, RL has not only been amenable to logical problems but human problems as well. Here is where we begin questioning and comparing the optimallity of solutions. Hence, in this paper we explore differing RL solution to a variation of human-based dilemmas (recent publications, 2017-on wards) and question whether the resulting models and frameworks appropriately and successfully capture the human problem. Notably, we focus on ethically and socially driven problems, many of which have been selected from the *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*.

Developing human-centric RL (HCRL) is not easy and often comes subject to the author's interpretation of the problem heuristic. Unfortunately, solutions typically lack meta-ethical understanding (where it exists), Miner, Petocz (2003). Consequently, HCRL requires comprehensive understanding and meticulous design to perform actions which are reasonably in line with human philosophies. This is imperative when faced with potentially harmful situations. We note this is a tribulation for research which relies on *world models*, such as safe autonomous driving. For example, Brunnbauer et al. (2021) investigates autonomous racing without any ethical safe-measures (i.e. what safe-measures exist if a manned autonomous racing car were to crash). On the other hand Cao et al. (2021), provides evidence for a "dangerous action policy", ensuring measures for driver safety. Accordingly, one could say the latter HCRL autonomous driving model has more consideration for a realistic world model, so is more likely reach the true optimal outcome, with respects both mathematical and meta-ethical interpretations of the problem. Therefore, our ambition to evaluate HCRL approaches is stimulated by a need for meta-ethical consideration, Miner, Petocz (2003).

This sentiment is echoed by Vouros (2022), who very recently (in the last month), published an outline of challenges and abstract resolutions for designing and implementing deep HCRL methods (they refer to this as "XDRL"). To the extent of our knowledge, there is no other research being done in assessing HCRL methods in this manner, therefore we feel it necessary to align our evaluations of existing systems with the research provided by Vouros (2022). Consequently, we structure this literature review by evaluating three key features of HCRL methods: *Transparency*, which is the ability a system has in generating explanations that are understandable in relation to actionable context (similar to the first principle defined in Rudin et al. (2022)), *Explainability*, which is the ability to provide qualitative understanding of response and objectives given the observed environment, and *Interprability*, which is a system's ability to further understand the content described by explainability.

Transparency, is a term Vouros (2022) concludes has become unfortunately synonymous with interprability. The two are linguistically distinguishable, yet the evaluation of infor-

[1] Context: https://www.theguardian.com/sport/2021/feb/12/deep-blue-computer-beats-kasparov-chess-1996

mation within a system has paired the two definitions. Therefore, our focus on transparency will be evaluating to which extent the current systems have applied pragmatic constraints for explanation. Thus we will come to ask whether particular systems employ methods for ensuring explanations are reasonable result of the environment.

Explainability is the easiest term to grasp, yet concluding remarks from Vouros (2022) indicate the difficulty associated with generalising methods of explanation over different levels of "scale and granularity"[2]. We will consequently evaluate the adaptability of explanation which current research provides us with, and criticise the design which may/may not exist for handling explanation in relation to potentially unforeseen events.

Our final term, interprability, may seem loaded with evaluating explanations however in many HCRL problems the way a system interprets the real world environment is strongly linked to the sequence of decisions it will come to make (this is particularly true for deep RL models). As we mentioned earlier RL can often evade the need for directly explaining models, however such thinking may also lead to questioning the fidelity of a system. If one does not know *why* a model should act a certain way, it may not be evident that it actually is acting that way - perhaps the model is acting in a similar yet somehow indistinguishable fashion. Hence, we outline the need for robust interprability.

In addition to evaluating these three properties of reasonable HCRL we make delineations on *AI alignment*. This discussions will be centred on the definition and evaluation of AI alignment as laid out by Butlin (2021). AI alignment refers to the promotion of reasonable human values, which directly effect the interpretation of state spaces as well as formulation of reward functions. The publication distinguishes the effects of promoting various human values and proposes topical forethought in relation to certain human objectives. For example, if we wanted to design a customer service robot with the objective of optimising financial return, one may design a simple reward function based on financial cost and return of service. In the short-term this may seem rewarding, however this neglects properties of customer service such as consumer satisfaction and probability of customer-returning. Thus such a design can be described as miss-aligned.

In the following sections, we will evaluate state-of-the-art models and discuss the results of each model publication in relation to pre-determined principles of responsible human-centric design. In Section 2 we will assess the contributions of Vouros (2022) and Butlin (2021) to define a set of principles for HCRL design. In Section 3 we will discuss several recent models and consequently evaluate their design in Section 4. In Section 5 we will draw conclusions and advise future design as a result.

2 Discussion

As we have previously mentioned, responsible design is necessary when proposing HCRL models, however deciding on what concepts can be considered responsible is rather difficult, Vouros (2022) and Butlin (2021). Consequently, we need to establish certain attributes a proposal needs to consider in order to qualify as "responsible", to which end we evaluate the propositions made by Vouros (2022) and Butlin (2021) and delineate our own principles of responsible implementation.

2.1 Interpretability, Explainability and Transparency

Vouros (2022) deliberates on key issues of interpretability, explainability and transparency, reasoning that consideration for ethics and trust needs to exist particularly in critical cases where both humans and agent play significant roles.

We can firstly relate interpretability to an agent's understanding of explanation for particular response. Momentarily disregarding the significance of explanation, an agents interpretation of rewards influences the formation of new policy, and visa versa. Thus, effective design for interpretability will not just evaluate the reward and policy models together but also individually. This ties in with aspirations for AI alignment in Butlin (2021) who detail the importance of pessimistic interpretations with psychologically complex tasks as a way of mediating successful completion of objective. Therefore, having some functionality focused on interpretability will ensure oversight in aligning values and policy is managed.

Interpretability is important, however perhaps not as important as explainability, which is a discriminating factor in resulting interpretation. Vouros (2022) distinguishes explanability as "qualitative understanding of results and observations". To implement this, a system should be able differentiate between what information should and should not be understood, constrained by the scale and granularity of observations. For example, if we provide a large and granular set of observations we can expect not everything needs to be explained - we don't want to over complicate our interpretation of explanations as it will overfit. One may want to simplify the implementation of explainability by monitoring a subset of features and evaluating their importance to aligned objectives. In doing so, the interpretation of explained features can help optimise policy, though more importantly by interpreting the significance of features we can determine if a system design is appropriately AI-aligned.

This method of evaluating significance is related to the transparency of a model. As mentioned previously, transparency is helpful in evaluating how pragmatic a set of explained features is to the alignment of our model. Hence, a transparent design will ensure explanations have relevance to the objective of our model.

[2] *Scale*: refers to the size of our problem. *Granularity* refers to the resolution of our observations

2.2 Alignment

Butlin (2021) argues that agents should be built to promote human-values. Drawing out the pros and cons of each set of values can be exhausting. Therefore, Butlin (2021) advises the characterisation of reward and environment to ensure values can be learned. In consequence the authors discuss many basic philosophies for appropriate behavioural learning.

The purpose of promoting human-values in this context can be misleading. You may reasonably assume that HCRL does not always require a behaviourally-oriented philosophical value system. The delineation we would like to make, is that human values do not have to be characterised by social philosophies, they can also be characterised by individual principle. Therefore, we extend the prior definition of AI alignment to encompass values which are not tied to emulating social expectation. For example, we don't need overbearing philosophical principles to teach a service robot manners. Hence, research does not need to rely on existing psychological literature to propose responsible HCRL models, as suggested by Butlin (2021). Yet, we do expect model proposals to provide extensive and reasoned explanation as to the design of inferred values, when such values are not directly explained by existing research.

In line with the prior section we expect to see these in explanations resulting from the reward function and world environment. Hence, model proposals should carry some method of meta-ethical inspection which provides evidence of embedded values.

2.3 Principles for Responsible Design

As a result of evaluating Vouros (2022) and Butlin (2021) we can define five principles which ensure responsible design. Note that we are not defining the only correct method, but outlining important principles surrounding ethical consideration and viability. Ensuring that responsible implementation follows considerations made towards a human-centric objective will be a primary criterion for model evaluation. The following principles only provide evidence of this, so we will be lenient towards publications which do not wholly encompass the principles. Though, some research negates the verification of embedded values (discussed in future sections); this is just bad practice and is not tolerated in our evaluations.

1. Assessment of reward and/or policy features should be *explainable*, *interpretable* and *transparent* in relation to direct design or potential consideration towards such design

2. Reward functions should hold real-world significance

3. Considerations should be made towards mediating unforeseen circumstances

4. Hence, if value system has been chosen it should cohere with such considerations

5. Research should provide viable ways of verifying embedded meta-ethical values

We reason (1) and (2) as a result of our prior evaluation. Principles (3) and (4) branch from our assessment of Butlin (2021), whereby we set the expectation that without successful mitigation of unforeseen events a model will not have embodied to correct values w.r.t to the real world environment. We acknowledge that this information may not be discussed in publications, however it is nonetheless useful in verifying the character of a system. Thus, Principle (5) enforces a similar sentiment in a more generalised sense.

Conclusively, we have evaluated current research to understand what is necessary in responsible design for HCRL systems. In the coming sections we will investigate whether current research takes these factors into account. We keenly note however, as made apparent in Vouros (2022), that considerations towards such layout of HCRL research is rare and often dependant on the interdisciplinary application. In mediating this expectation, we will advise areas for improvement.

2.4 Ethical Gravity Thesis for Automated Decision-Making

Before concluding this chapter we would like to discuss the importance of hierarchical considerations made towards differing levels of problem abstraction. Considering the process for defining levels of problem representation is not one which often makes it into the literature, we have chosen to ignore it in our evaluations of novel research. Nonetheless, we will evaluate recent literature which explores this process to determine its value in responsible design.

Automating the decision-making processes is difficult for several reasons. Firstly, the definitions which construct the abstract representation of ethical dilemmas are hard to comprehend. Secondly, implementation-level abstractions can be difficult to define. Thirdly, ethical problems which occur at high-levels of abstraction may affect lower-level of computation within an automated system. The *Ethical Gravity Thesis* (EGT) defends the last reason, stating that problems which arise at"higher levels of analysis...are inherited by lower levels of analysis ", Kasirzadeh, Klein (2021). For example, the design for algorithmic-level analysis (i.e. explaining the algorithmic representation of the problem) will directly influence implementation-level explainability.

Thus, the expectation for solving abstraction-level bias is presented in Kasirzadeh, Klein (2021). Whereby, Marr's 1982 framework, Kitcher (1988) is used to define the levels of abstractions necessary to exemplify aspects of responsible design considerations. Marrian hierarchy defines four levels of analysis, (1) Functional-level analysis, which considers the functional objective of the entire system, (2) Computational-level analysis, which encompasses the abstract computations necessary to carry out functional analysis, (3) Algorithmic-level analysis, which defines the sequence of procedures necessary to perform computation, and (4) Implementations-level analysis which directly explains how the algorithm will be implemented to attain the objective through automation. We recommend Figure 1 Kasirzadeh, Klein (2021) for visual representation of this hierarchy.

Consequent to defining this framework, Kasirzadeh, Klein (2021) defends EGT through two points of reasoning, (i) the *Realisation Argument*, (ii) the *Institutional Argument*. The Realisation Argument states that lower-levels of analysis should realise the objectives of higher level of analysis however, higher levels of analysis do not necessarily require realising lower levels. Given this relationship for realisation, EGT makes sense, considering erroneously defined high-level objectives will result in (at the lowest-level) implementations problems (can alternatively exists as ethical problems in implementation design). The Institutional Argument delineates from the process of creating and maintaining automated systems and describes the issues one is bound to encounter as a result of institutional constraints in developing automated systems. For example, projects are held accountable to financial and application incentives within cooperative structures, if the functional and computational level designs are shifted to cohere with cooperative interest then the problem will reach partial realisation. As we know from extensive technological (capitalistic) progress, this is computationally acceptable however Kasirzadeh, Klein (2021) argues this action is irresponsible w.r.t EGT and will certainly lead to ethical problems. We have seen many examples of these in the past decade, the most infamous being the *Cambridge Analytica* scandal surrounding unfair use and distribution of private-data at implementation level as a result of poorly defined computational level design.

Kasirzadeh, Klein (2021) proposes interesting forethought to design. Not only do there exist benefits to EGT in the ethical sense, but one will often find the hierarchical structure coheres, and can be mapped, to a work-flow process for design. In addition, it is reasonable and logical to assume that without a relationship of realisation between the highest and lowest states, no objectives can be ethically realised.

In reviewing Kasirzadeh, Klein (2021) we come to the conclusion that such processes are necessary for ethical design however we are not suggesting that Mar's 1982 framework exemplifies an ideal solution. Considering that human-centric automation will, more often than not, require abstract solutions, resulting complexities may be difficult (if not impossible) to solve at higher-levels. Thus, lower levels of analysis will consequently suffer, as per EGT. We can mediate these problems by re-defining the highest level objectives as a set of multi-objective problem so that consequent implementation (or lower) levels, through the realisation argument, will only be affected by directly related objectives. The hierarchical structure for this suggestion is illustrated in Figure 1.

Considering the abstraction-level design is not discussed in literature and is often interchangeably used as a work-flow process for defining the problem prior to defining the solution, we can neglect it in the following evaluations. Nonetheless, we would find it irregular for HCRL not to include some forethought to levels of problem representation.

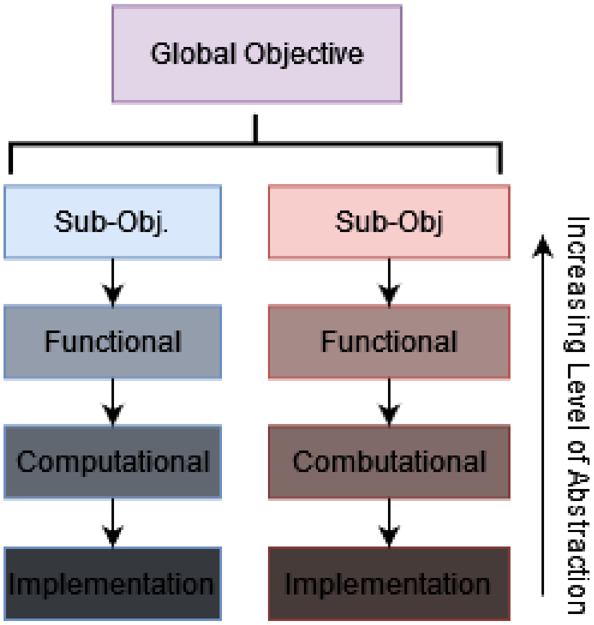

Fig. 1. Marrian Hierarchy of a Multi-Objective human-centric problem, where global objectives are redefined as sub-objectives, which define the highest level of abstraction

3 State of Art

In this section we present recently published HCRL models which will be evaluated in Section 4. The purpose of this section is not to explain the technical details of each model, but to provide insight into particular choices made for design or testing which we feel distinguishes the proposed research from competing research. Thus, each paper will be presented with a brief overview of the model's purpose followed by a discussion outlining the prominent features of the model.

3.1 Influencing Multiple Agents for Exploration

A classical example of a difficult-to-solve HCRL problem, is the mediation of interactions in multi-agent systems. This is relevant to spare-reward tasks - often the case real-world environments. Particularly, rather than being dependant on a totally collaborative target as one may initially think, behavioural values depend more on intrinsic (individual) curiosity; models which optimise this commonly characterise said behaviours through additional incentives which boost exploration, such as *pseudo-counts*, Bellemare et al. (2016). However, often this comes at the cost of a single-agent system, which is evidently not sufficient for a realistic world-environment. This problem more often than not, expects multi-agent cooperation.

Wang et al. (2019) looks at solving this by introducing two methods; each criticise different aspects of problem thus contain different optimisation objectives. The first method is *exploration via information-theoretic influence* (EITI), which encourages one agent to explore known critical positions to better influence the transition probability of a different agent, optimising mutually shared information. The second method is *exploration via decision-theoretic influence* (EDTI), which builds on EITI by separating transition and reward influences by measuring the *Value of Interaction* between agents. Thus, EDTI encourages coordinated exploration while filtering out meaningless interactions which do not lead to global coordinated objective.

3.2 Analysing Cooperation for Multi-Agent Sequential Social Dilemmas

Learnable scenarios for multi-agent social games are difficult to find yet more difficult to design. However, analysing behavioural aspects of cooperation under conditions such as individual power of choice to behave greedily/altruistically has been consistently researched. Often, we find systems that analyse behavioural aspects of socially formed cooperation take the structure of a genetic algorithm (GA), though recently RL methods have shown to provide additional benefit to simulating social scenarios. While it is true that the computational cost of RL compared to GAs means RL has a lower hosting capacity (RL simulation involves significantly less individuals), GA does not easily provide us with the capability of solving complex scenarios which pit greed and cooperation.

Matrix game social dilemma (MGSD) grid worlds procure such scenarios, thus can be used to evaluate the behavioural dynamics of multi-agent systems[3]. Leibo et al. (2017) evaluates MGSDs and criticises several real-world characteristics which are ignored. The characteristics of importance are, (1) Cooperation can be a graded quantity (where MGSDs treat cooperation as an action rather than a policy), (2) individual decisions can simultaneously influence and be influenced by other agents, and (3) individuals maintain partial-observability of environment and other players.

3.3 Training to Optimise Aspects of Moral Uncertainty

Deploying autonomous agents, take our example of autonomous vehicles in Section 1, requires safety measures, not simply focused on physical safety but ethical and moral safety for the customers and product owner of the vehicle as well as the engineers who designed the vehicle. This problem is constrained by characteristics which promote values of moral uncertainty. We can refer back to the prior subsection for an example, where the relationship between altruism and selfishness was assessed. Situations analogous to “safe-crashing” will inflict the most uncertain moral dilemmas, thus solving the problem of moral uncertainty is imperative.

In doing so, Ecoffet, Lehman (2021) evaluates training methods to mediate morally uncertain circumstances. In philosophical terms, Ecoffet, Lehman (2021) outlines the difficulty of the problem by comparing *Utilitarian* decision making to *Deontological* decision making; the stark difference being that utilitarian philosophy forgives rule-breaking for global good while deontological philosophy follows absolute law. The question then becomes, when and how should a utilitarian (or similar) approach be considered?

Consequently, Ecoffet, Lehman (2021) criticises scaling worthiness of choice derived from the sum of a sequential set of utility values for each philosophical theory, as scaling is sensitive to the scope of observed scenario (think of the difference between utilitarian response for global good vs local good w.r.t safe-crashing). Instead, they suggest a scaled voting-based system which invokes a principle of “Proportional Say”, as a method of training a policy based on a set of coupled philosophical theories. The voting is accomplished using expert hand-labelling of datat. Hence, the influence of choice relies on the credibility of a theory rather than the maximising the worthiness of sequential choices.

Ecoffet, Lehman (2021) concludes that some ethical theories are fundamentally incomparable thus scaling their choice of worthiness is futile. The alternative voting-system, though viable is still difficult to solve and relies on evaluations made by experts. As a result they suggest an investigation of varying theoretical desireta for more complex social dilemmas - dilemmas which infer realistically relatable characteristics.

3.4 Learning Socially Acceptable Behaviours for Robot Waiters

Humanoid Service Robots (HSRs) observe customer’s visual and verbal reactions to better tailor the dining experience. This suggests variations in the speed of a robot-waiter, the direction of movements, physical/verbal responses to a customer and areas of movement. McQuillin et al. (2022) provides a solution to this “problem” by proposing a crowsourced labeled custom data set for pre-training, an online method for continuous training and explicit and implicit reward mechanisms to allow for adaptation towards a customers desires for better service.

With regards to model design, McQuillin et al. (2022) proposes various components for behaviour-related adaptation such as automatic facial reaction analysis. The aim of the proposed model is to evaluate the results of explicit feedback (derived from the crowd-sourced data), implicit feedback (derived from the customer) and a mechanism for combining implicit and explicit feedback. The data set which McQuillin et al. (2022) provides is based on a sample of students studying at the University of Cambridge, the majority of which are “members of the Computer Science department”. A similar set up exists for the testing of the model, whereby 8-11 “unique” participants were served by the robot for model evaluation purposes. This was done by assessing the speed, visibility and movement of the robo-waiter, and matching this action to a 2-D map of the environment where individuals assessed the areas they felt most/least comfortable with respect to the criteria.

The paper concludes by evidencing the betterment of explicit data over the use of implicit data, suggesting that pre-

[3] a matrix game becomes a social dilemma when cooperators and non-cooperators can be influenced into defecting from their chosen behavioural policy

training a robot rather than inciting an online-form of training results in a more enjoyable and sociable experience. The paper outlines that due to COVID-19 limitations in diversity do exist in the data set, yet they state it is still useful in generating a "good" experience for customers.

3.5 Conditional Generative Adversarial Imitation Learning for Taxi-services

The problem of designing a service application using region specific data will bias the resulting service, particularly when successful deployment of an application is sensitive to regional characteristic. This is a prominent problem in the assisted-search of customers for idle taxis. Knowing where to go, when to go and the optimal path to get to a potential customer is a quality one will find in only the most experience taxi drivers. In addition, the preference of local region which a taxi driver may want to explore for eager customers is not often thought of when designing such solutions. Thus, responsible solutions which aid customer-search are difficult to find.

Zhang et al. (2019) provides a solution to this problem by proposing a conditional generative adversarial imitation learning (cGAIL) model which relies on driver-specific preferences and the trajectory and consequent feedback of trajectories from drivers around specified locations. Hence, the movement of other drivers and qualitative feedback of experience searching for customers in a designated region are taken into account when determining optimal trajectory.

Through the use of imitation learning, Zhang et al. (2019) promises a personalised experience record and preferred trajectory as a result of the user's life-style and preference choices. Training is accomplished by exploring options outside a users specified preferences (baring preferences such as location of residence, working hours, etc.) and evaluating the reward of exploring such regions. Zhang et al. (2019) delineates that knowledge learnt is (1) transferable across agents and (2) transferable across trajectories (tasks with similar objectives and parameters in differing regions expect similar policies for training). This is evidenced in the results section of the publication, where authors took three months of data from Shenzhen, China, and found the cGAIL model not only demonstrated learned preference but achieved a significant margin of success over other state-of-there-art baseline approaches (quoted as 34% more accurate).

Similarly to the prior section a sample of $\sim 11,000$ drives was taken, $3,000$ of which were qualified as "experts". As a result the model was informed (pre-trained) using their professional insight to improve performance in practice.

3.6 Hyper-Meta Learning for Sparse Rewards

We have outlined previously the difficulty with sparse-rewards, yet it holds great significance when training HCRL models responsibly. An option for solving such problem is to use hyper-meta reinforcement learning (HMRL), proposed in Hua et al. (2021), which "extracts meta knowledge to help efficient policy learning on new tasks". In this manner, it presents a method for generalisation by learning from meta data rather than pure data.

The HMRL model primarily consists of a "cross-environment meta state embedding" module which abstracts the state-space by extracting meta-data relating to an environment and compares this with a previously learned similar meta state space to reduce the impact of negative influence of spare rewards. Hence, the meta reward shaping technique proposed is fed the meta knowledge learned from current and other environments. In technical terms, the meta knowledge learned can result in intrinsic reward as a result of other meta state-spaces, reducing the impact of expected sparse reward.

In evaluating this novel tool, Hua et al. (2021) compares HMLR with four state-of-the-art baseline models for fully-observed and partially-observed visual state-spaces. We note that the description of this tool is rather convoluted; to better understand how meta knowledge is evaluated, the value-maps which result from reward-shaping tool can be visualised as heat maps, denoting regions of interest or exploitation (refer to Figure 5 Haykin, Network (2004) to see how three varying environment set-ups for a small maze procure similar information of interest through the use of heat-maps illustrating regions of potential value). The paper triumphs over the state-of-the-art, and evidences superiority in transfer-ability of meta knowledge and efficiency of training.

4 Literature Analysis

In this section we analyse the papers presented in the previous section with regards to our principles for responsible design. Despite the incomparable differences between each presented model, we can evaluate the responsibility of design with respect to one another to orm an impression as to what trends may exists within HCRL.

Most prominently, we found trends in reward-function design considerations to ensure appropriate emulation of realistic environments, which is hopeful for the progression of HCRL which has traditionally suffered from unrealistic environment simulation. On the other hand, we still found a majority of literature did not test for the complex (more realistic) representation of presented problems but for simplified versions, such as simulating social interaction with *Prisoners Dilemma* or *Wolfpack Game* as opposed to testing in actual practice. Conclusively, we look forward to the advancement of simulating realistic environments as the interest in redefining reward function progresses.

4.1 Influencing Multiple Agents for Exploration

EITI looks at influence as a result of the entire environment which may be more behaviourally responsible if one is uncertain as to what different interactions can lead to. In an environment where less information is initially known, (particularly regarding HCRL problem which we use for understanding social functionality in an unknown environment through emulation rather than problem solving), discounting interactions earlier on can tend to instability later in training. Hence, EITI is a relatively responsible method of multi-agent exploration for more realistic world-environments. However, EDTI still provides us with a method for optimisation when we do completely understand which transitions can be negated as a result of poor future extrinsic reward. We see in the analysis section of Wang et al. (2019) (which tests simple 2-agent, N-agent and large-scale problems) that both EDTI and EITI provide benefit, where EITI is the dominant solution when points of interaction (for example global positions in a maze) directly line up with extrinsic reward, and EDTI is the dominant solution when this is not the case as it is able to filter out interaction points which are not so beneficial.

With respect to the principles for responsible design, Wang et al. (2019) provides us with two realistic reward functions with evident consideration made towards mediating value systems. To this extent, behaviours from EITI and EDTI value systems are verified within the paper, which suggests that Wang et al. (2019) abides by our principles of responsible design.

Though the paper does not provide a method for explainable, thus interpretable and transparent, assessment of features it perhaps is not absolutely necessary in the context of the paper - which is to encourage successful cooperation in a less abstract scenario. If one were to utilise EITI (which we deem as the more responsible method for more abstract HCRL multi-agent problems) we would expect a level of testing which analyses the embodiment of cooperation, perhaps less by evaluating quality of interaction-types and subsequent state transitions and more by evaluating the abstract human-value of such interactions. This would be a more realistic framework for *continuous* world-environments.

4.2 Analysing Cooperation for Multi-Agent Sequential Social Dilemmas

It is evident that the proposed SSDs take a very responsible approach to re-evaluating the context of social-dilemma games. The critical characteristics which were missing from MGSDs are extended and show benefit. Though this does not suggest the entire problem has been captured.

We criticise characteristic (2) in Section 3.2 as the simultaneity of a pair's actions can only exist if, (i) the influence from a cooperative actor (if a large social cooperative group is formed) is greater than the influence of a selfish actor, and (ii) agents need to first locally interact before influencing one another. It may seem counter intuitive to expect that cooperative power holds greater power, however we need to separate our understanding of group-influence from selfish policy - meaning that groups of selfish individuals do not have collective influence while cooperative individuals do. This is the indirect appeal of cooperation (the direct appeal being pragmatic use of resources). To this degree we suggest that mechanisms for mediating influence needs to be more responsibly designed, whether this is in the manner we have outlined or similar.

In evaluating the principles we laid out in Section 2 Leibo et al. (2017) provides an apt assessment of the policy features and explainable, interpretable and transparent design[4]. Evidently, no considerations are made to reward-functions however this may not be relevant to the paper as it explores the construction of realistic environments. Overall, the paper shows promise towards accreditation and validation of realistic environment-spaces.

4.3 Training to Optimise Aspects of Moral Uncertainty

It is clear from Ecoffet, Lehman (2021) that there exists promise in voter-based systems and, as with prior assessment, takes a step in the right direction w.r.t to finding a final solution. However, the results of this paper only suggests future steps toward voting-systems, which already have been previously considered to further extent, as outlined in Azzarelli (2021).

With respect to the our principles for responsible design, what is clear is that there exists extensive consideration for mediating unforeseen events with evidence of an effective value system (in this case the value system is voter-based). Despite this we have to criticise the implication of a voter based system which relies on expert labeled-data. As outlined by Ecoffet, Lehman (2021), moral uncertainty is a rather new research field. Thus, our expectation for professional insight will be relatively poor[5] as solid theory and practice has not been determined. The consequence of this is a rather large error in human labeled-data which will be the minimum achievable error of any RL model. Conclusively, we suggest that more research needs to be done to reduce this error and if this is not achievable, a solution to the unsupervised RL problem needs to be investigated.

4.4 Learning Socially Acceptable Behaviours for Robo-Waiters

In evaluating McQuillin et al. (2022) we have to acknowledge the difficulty of testing such a model while under UK COVID-19 restrictions. In spite of this, we cannot accept hurrying of research for publication-sake. So, with regards to the article we have to wonder what the intentions were in testing and publishing in this fashion.

Although the evidence of irresponsible design is almost unquestionable, we can assume that the trained model will perform well in local areas with a similar social-infrastructure to the Cambridge-data. Therefore, we aren't necessarily questioning the utility, but the capability for general application of this model which has been inferred from the publication.

The final suggestion we would have for papers which want to explore service robots in a similar manner is the use of expert-understanding when it comes to waiting-etiquette; successful examples of this have been shown in Zhang et al.

[4] This isn't directly discussed but considerations have been made

[5] Perhaps not completely true for simple situations

(2019) and Ecoffet, Lehman (2021). The etiquette is dependant on the type of food-service you are providing, for example fast food traditionally optimises time-related costs while Michelin-star restaurant will optimise the elegance experience.

4.5 Conditional Generative Adversarial Imitation Learning for Taxi-services

As with the prior evaluation we could criticise Zhang et al. (2019) for testing cGAIL on a singular instance (reviewing the application over one region), however we believe the ambition of the two papers is made different. While McQuillin et al. (2022) aims to generalise the service of a singular waiter for a room full of customers, Zhang et al. (2019) looks at specialising the service to one user and attentively mediating target trajectories by assessing the intentions and history of other users. Thus, our previous criticism is less appropriate in this case.

What Zhang et al. (2019) demonstrates is the ability to consider the experience and expectation of a driver to boost financial return. Through explicit formation of both policy and reward networks, the model is capable of optimising several parameters pertaining to driving experience. The publication provides the formation of these networks in an explicit and transparent way through clear analysis of mathematical definitions and consequent algorithms. The resulting application is able to consider not just unforeseen events but adapt individually through the learning of non-linear reward policies (the publication highlights this as a reason why current state-of-the-art doesn't perform so well - their policy networks are linearly dependant on a reward function).

4.6 Hyper-Meta Learning for Sparse Rewards

From the range of designs we have evaluated this is the only paper which directly embeds explainable, interpretable and transparent components within design. To this extent, it is able to successfully explain and interpret result during learning to shape the meta reward function and update the meta policy. As well as this, the model is able to transfer explainable criteria between environments through the learning of meta state-spaces. What is particularly interesting is the joined gain in efficiency and understanding this model has over state-of-the-art; where one would usually find compromise between aspects of efficiency and performance.

We find it difficult to flaw the design of this model as it coheres explicitly with all the principles of responsible design.

5 Concluding Remarks

HCRL is a problem-field which imposes a complex sum of abstract constraints, many which are not so easily understood. The consequence of this can lead to mis-representation and bad practice when presenting novel design in research. On the other hand, it provides room for exploration and creativity; resulting solutions will seem more impressive.

In reviewing a (small) sample of recently published HCRL papers, we have provided examples which land between these two extremes. In addition, we criticised the literature and consequent models against a set of well defined criteria influenced by two additional (recent) publications which we had previously evaluated for use.

As a result of the evaluation, we can delineate several points of interest for future research. The first point is that we have realised a steady trend in the number of publications oriented towards improving the design of the reward function. We note there is additional interest towards solving spare-reward tasks. Despite the evident benefits, we should address the worrying level consideration made towards direct design for explainable, interpretable and transparent criteria. As we discussed in Section 2, this hierarchical logic is necessary to ensure understandable and evident success of HCRL models. Without consideration towards such design, resulting models will be less performative in actual practice. In evidence of this, we found the more successful models (HMRL and cGAIL) satisfied such criteria and consequent results were more verifiable.

The second point is oriented towards progress which needs to be made in certain philosophical fields, such as meta-ethics and decision making ethics for scenarios which are socially/morally comparable. Without such understanding, presenting badly informed research will result in skewed results as was the case with the robo-waiter. We can relate this to poor understanding of EGT, whereby solutions with badly defined abstract problem representations will suffer on the implementation level (hence, the argument for responsible realisation of abstract levels). Thus, in Section 2 we suggested a new layout of abstract design structure which could solve parts of the problem. Despite this, considerably more effort still needs to be made towards outlining the philosophical/psychological nature of the problem. When this becomes difficult to evaluate one should consider seeking counsel from professional within that discipline, as was done in Ecoffet, Lehman (2021) and Zhang et al. (2019). In this manner, we would not have to utilise the mutli-objective Marrian hierarchy, which would be significantly more difficult to outline than the simplified hierarchy presented in Kasirzadeh, Klein (2021).

In finality, we look forward to the progression of the interdisciplinary HCRL research, with particular interest towards the exploration of meta-learning. This concept adds an interesting characteristic to input data and extends RL models to a higher order of utility. With the current proposal for HMRL we are intrigued to see how this design will be extended. Additionally, we hope to see further research oriented towards testing realistic social dilemmas rather than simulating the characteristics of such dilemmas through simple games. We anticipate a new set of baseline scenarios which will hopefully advance HCRL towards solving more abstract tasks.

References

Azzarelli Adrian. A Framework for a Decision-Making Robot regarding Generalised Ethical Dilemmas using ML // University of Southampton UG Dissertation. 2021.

Bellemare Marc, Srinivasan Sriram, Ostrovski Georg, Schaul Tom, Saxton David, Munos Remi. Unifying count-based exploration and intrinsic motivation // Advances in neural information processing systems. 2016. 29.

Brunnbauer Axel, Berducci Luigi, Brandstätter Andreas, Lechner Mathias, Hasani Ramin, Rus Daniela, Grosu Radu. Model-based versus model-free deep reinforcement learning for autonomous racing cars // arXiv preprint arXiv:2103.04909. 2021.

Butlin Patrick. AI Alignment and Human Reward // Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021. 437–445.

Cao Zhong, Xu Shaobing, Peng Huei, Yang Diange, Zidek Robert. Confidence-aware reinforcement learning for self-driving cars // IEEE Transactions on Intelligent Transportation Systems. 2021.

Ecoffet Adrien, Lehman Joel. Reinforcement learning under moral uncertainty // International Conference on Machine Learning. 2021. 2926–2936.

Haykin Simon, Network N. A comprehensive foundation // Neural networks. 2004. 2, 2004. 41.

Hua Yun, Wang Xiangfeng, Jin Bo, Li Wenhao, Yan Junchi, He Xiaofeng, Zha Hongyuan. HMRL: Hyper-Meta Learning for Sparse Reward Reinforcement Learning Problem // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. 637–645.

Kasirzadeh Atoosa, Klein Colin. The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-making Systems // Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021. 618–626.

Kitcher Patricia. Marr's computational theory of vision // Philosophy of Science. 1988. 55, 1. 1–24.

Leibo Joel Z, Zambaldi Vinicius, Lanctot Marc, Marecki Janusz, Graepel Thore. Multi-agent reinforcement learning in sequential social dilemmas // arXiv preprint arXiv:1702.03037. 2017.

McQuillin Emily, Churamani Nikhil, Gunes Hatice. Learning socially appropriate robo-waiter behaviours through real-time user feedback // Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction. 2022. 541–550.

Miner Maureen, Petocz Agnes. Moral theory in ethical decision making: Problems, clarifications and recommendations from a psychological perspective // Journal of Business ethics. 2003. 42, 1. 11–25.

Rudin Cynthia, Chen Chaofan, Chen Zhi, Huang Haiyang, Semenova Lesia, Zhong Chudi. Interpretable machine learning: Fundamental principles and 10 grand challenges // Statistics Surveys. 2022. 16. 1–85.

Silver David, Huang Aja, Maddison Chris J, Guez Arthur, Sifre Laurent, Van Den Driessche George, Schrittwieser Julian, Antonoglou Ioannis, Panneershelvam Veda, Lanctot Marc, others . Mastering the game of Go with deep neural networks and tree search // nature. 2016. 529, 7587. 484–489.

Vouros George A. Explainable Deep Reinforcement Learning: State of the Art and Challenges // ACM Computing Surveys (CSUR). 2022.

Wang Tonghan, Wang Jianhao, Wu Yi, Zhang Chongjie. Influence-based multi-agent exploration // arXiv preprint arXiv:1910.05512. 2019.

Zhang Xin, Li Yanhua, Zhou Xun, Luo Jun. Unveiling taxi drivers' strategies via cgail: Conditional generative adversarial imitation learning // 2019 IEEE International Conference on Data Mining (ICDM). 2019. 1480–1485.

Word Counts

This section is *not* included in the word count.

Notes on Nature Methods Brief Communication

- Abstract: 3 sentences, 70 words.
- Main text: 3 pages, 2 figures, 1000-1500 words, more figures possible if under 3 pages