Research Statement

Reid McIlroy-Young

In 1956 the seminal Dartmouth Workshop formalized the goals of Artificial Intelligence (AI) research: language understanding, learning algorithms, self improvement, abstraction, interpreting neural networks, etc. In the past half century some of these goals have been achieved. We have learning algorithms that can surpass human skill level on a variety of tasks, from language comprehension, to image recognition and game playing. Some of these AI systems are already embedded in the real world: tens of trillions of dollars in commerce a year are overseen by fraud prevention algorithms, facial recognition systems are available in virtually every phone and camera on the market, and recommendations from AI systems are a common way people encounter new ideas and products. However, other areas such as the understanding of how these systems make decisions have had limited progress and self-improvement continues to elude the grasp of researchers. Deep neural networks use "feature generation", a type of abstraction whose interpretation is still beyond our ken in non-trivial settings. Because AI systems are capable of making decisions or judgments on the real world, interpretability of these models has arisen as an important goal. Interpretability is needed because it is fundamental for certain applications to understand the mechanisms that lead to a decision, and because it is desirable that these systems have additional utility beyond the maximisation of performance on a highly prescribed task. This need is particularly strong when an AI system exceeds human performance (*i.e.*, is *superhuman*).

How can we bridge the gap between artificial intelligence and human intelligence to build mutual understanding between the two paradigms of knowledge?. In my research, I have studied and built systems where collaboration between humans and AI can contribute to answering this question. My work looks at activities that both human and AI undertake, and in particular activities where computational methods are superhuman. I analyze how the AI and human approaches differ, with the ultimate goal of building ways for humans to learn from their computational compatriots.

The system of focus for my work is chess. Chess has a long history of intersections with AI, from the mechanical Turk, to Turing's work on chess algorithms, Deep Blue and AlphaZero. All these algorithms were at the forefront of developments in the field. Chess is a model system with attractive properties for AI research: large datasets of diverse players are easily available, as are superhuman engines, both AI and non-AI. A final property is that chess is benign, i.e., the potential negative consequences of research into it on society are minimal. These factors led to my main research project, building a new style of AI chess engines, ones that attempt to mimic human play—instead of ones that achieve higher performance. The models, called *Maia*[1], that I released as part of this work have been well received: 2+ million games played against the versions I run, incorporated into commercially produced chess systems and even written about in a textbook. My models are now the baseline for what it means for an engine to be human-like in chess. The challenge of this work was not creating the models, but it was in properly formulating the question 'What is human-like behaviour?' into a quantifiable form. This work shows that by working to build bridges between AI and human cognition we can both create teachers for humans and improve our understanding of computational intelligence.

Below, I outline the main research I undertook as part of my PhD, how I have built collaborative AI systems and how I have used my work to contribute to the AI research community.

From Superhuman to Human

There are few areas where AI is unquestionably superhuman, that is, when a AI system exceeds human capability on a task for all requests, even those far outside the scenarios it was trained in. Generalization, however, is the Achilles's heel of most attempts at superhuman AI. AI systems that can respond to previously unseen scenarios are desirable since they can be used to support human activity,

such as in the case of autonomous vehicles, high-precision surgeries, and human teaching. An example of the latter is AI in chess. The AI chess community has seen numerous AI systems over the last decades, from Deep Blue to more recent ones with superhuman characteristics. However, despite having built these superhuman AI systems, we humans struggle to learn from them. An early motivation of my project was thus to bridge this gap, by creating an AI model that is derived from a superhuman AI but it is understandable by humans.

Human-like Chess AI. There are of course many chess engines, which are programs that play chess, and can regularly lose to humans. So, one could have the impression that if these engines are defeatable, they already have human-compatible behaviour. In my work[1], I analysed the behavior of non-AI (*Stockfish*) and AI-based systems (*Leela* an *AlphaZero* implementation) when their win rates are closer to that of an average player. This is done by increasing the randomness of its actions, restricting computing resources, or limiting training data. The result is a model that plays a combination of strong moves and unpredictable moves, thus, not something that plays like a human. Players reported that *Leela* behaved closer to a human when compared to *Stockfish*, but they could still see that their opponent was not human. Thus, I proposed for the first time a way of quantifying how human-like the models are beyond a simple measure of performance.

A first attempt at measuring humanness could be to take a move and see how often humans make that same move. This does not work due to the 'combinatorial explosion' of possible positions in a game; after 5 moves by both players +50% of games of chess are in a position that has never been seen before by a human[4]. The measure I developed uses a large number of human games to measure the model's performance relative to humans directly at the move level. This measure that I called *move matching accuracy* thus provides a direct measure of humanness and has since been adapted to measure humanness in other games. Since players of different skill levels must be making different decisions, I also used the *move matching accuracy* to examine how human the models are across many player skill levels.

Once I translated my research objective into a quantitative framework, I was able to start developing my own deep learning chess AI system. My explorations led to an interesting discovery: reinforcement learning underperformed on move matching accuracy compared to simpler prediction-based models. As a result, my final model, *Maia*[1], is a deep neural network that directly predicts the actions a player will take on a given chess board, without using the typical reinforcement learning search. The *Maia* model is also tunable, which means I can accurately model players of different skill levels. When I compared *Maia* to the other engines *Stockfish* and *Leela*, it outperformed both by over 5 percentage points across all player skill levels. Interestingly, the main contributor to *Maia*'s performance is its ability to predict the mistakes humans make. While *Stockfish* and *Leela* are accurate when the target is playing well, *Maia* predicts minor errors with over 60 accuracy. This work was the first of its kind, showing that superhuman systems are different from humans in fundamental ways. *Maia* has had a significant impact on the chess community, even getting the attention of Garry Kasparov.

Individualized Chess AI. Building a human-like chess AI was the first step towards creating a system that can act as a teacher. To build on this work I created a method for *individualizing* the models, that is, to take our models that predict average human play and convert them to ones that predict the actions of a specific player. I first demonstrated that these *Individualized Maia*[4]models have higher *move matching accuracy* on their targeted players than any other model version. I also showed that this increase comes from predicting the minor errors that players make — the models learn where the player is imperfect. Robustness of the new models was a concern, as they may just be memorizing a few common patterns that the players employ, so I proved that the *Individualized Maia* outperformed the normal *Maia* models when given at game states where the targeted player was outside their normal play patterns. When I looked towards the broader goal of building an algorithmic teacher, these models allow me to analyse a player's games and not just find the mistakes, but highlight the learning opportunities by identifying

which errors are typical to that specific player.

These models also demonstrated an interesting property: their accuracy in identifying players they were not targeting is lower. This means that with only a small number of chess games, the models can deanonymize their individualized player from a set of candidates. This type of purely behavioral stylistic identification, which is called "stylometry," had not been demonstrated before, and I explored it further in my next papers[2,3]. I believe this work overlaps greatly with Dr. Brian McFee's work on music recommendations.

Intersections of Human and AI Intelligence

In my work using chess as a model system I have sought to formalize my work into concepts that apply to the broader AI field. My work on creating individualized behavioral models inspired to two different research publications that introduced novel concepts and techniques.

Behavioral Stylometry. To build an AI teaching system we will need an AI system that can understand people as individuals, learn what someone's particular weaknesses and strengths are. My work on individualized behavioral models showed that with only a few example chess games, a player can be be identified based on their style. Therefore, an algorithm should be able to understand someone's style with only a small number of games. With this insight, I created a deep learning (transformer and ResNet) based system that can take in an arbitrary number of games by a player and create a "player vector" that encodes their play style. These vectors allow for player identification from even a large number of other players, I achieved a 98% identification rate using 100 games from a set of 2800 candidate players, and when excluding the openings still achieved 86% accuracy showing that the model is learning the play style not just memorizing the players' favourite opening patterns. This work presents a first step toward creating an algorithmic teacher, since we want to create a teacher that can accurately understand the subtleties of its students. This work also revealed that purely behavioral data can be used for the identification of people, a novel privacy concern leading to coverage in *Science*.

Ethics of Modeling Specific People. When I started discussing how to release our code and models for the *individualized Maias*[4]paper with my coauthors we realized that there was no discussion in the literature as to the ethical implications of releasing these models. So, we worked with an ethicist to write a paper[3]that first defines what we are concerned about. We created the new term *mimetic model* as a model that simulates the decisions and behavior of a specific person in a given domain, *e.g.* modeling what an artist will do with a given prompt or what someone will write in response to an email. After defining and situating the term within the literature, there are many related concepts such as deepfakes, recommendation systems and style transfer that required discussion. We then described the many ethical concerns that the usage of these models may, or already are, causing.

To explore these ethical issues, we divided the types of concern into two categories. The first was model usage as a means to an end, such as modeling someone who is about to interview you for a job in order to manipulate and deceive them into giving you the job. The second was the usage of *Mimetic Models* as an end in themselves, such as training a model of a salesperson to respond to their emails when they are unavailable, which could devalue them and potentially mislead their customers. We also discussed the possibility of using models to model oneself, which raises additional concerns, such as whether artists who generate copies of their own work have the same value as an 'original'. In the months after the papers release, it has been interesting to see that our discussions of the implications of mimetic models on the arts have become a mainstream concern due to the release of text-to-image models like Stable Diffusion, and I plan to continue to be on the leading edge of human-centered AI.

Research Agenda

In my research, I seek to both answer questions about humans using AI and the converse, understand better AI using observations of humans. As such, I am interested in finding areas where there is overlap in the goals of both human and AI systems; availability of large amounts of behavioural data; and where the results can be extended to other domains. I believe we are at the start of a paradigm shift in AI, leading to opportunities for this type of research. As has already happened in chess will happen across other areas. Just this year for example, we have seen the start of AI collaborations with humans in writing software, creating art, and developing medicine. Developing a foundation for creating AI systems that function alongside humans will make integrating these new technologies into everyday life easier, the resulting AI systems more trustworthy and the AI systems performance more accurate. Below, I describe some of the approaches to these goals that I am excited to engage with.

Behavioral Modelling Beyond Chess. In my work on chess AI[1,4], I demonstrated that modeling human behavior requires a different approach than maximizing performance. Specifically, I showed that the difficulty lies in framing the task in such a way that machine learning techniques can be applied. Expanding and formalizing these techniques to other tasks has always been my goal, and I have already applied my work to other domains at FAIR, such as Go and Hex. I plan to continue branching out to other domains, particularly in modeling students learning algebra, programmers of different skill levels, and how clerks use their job-specific computer systems. By creating AI systems that model humans, the designers of the systems will be able to make them better collaborators with their users.

Ethics and Human-like AI. In my work I have shown that making AI systems that mimic humans[4,3] or understand human's style[2] raises novel ethical concerns. By working with a strong model system we can test models that are too computationally intensive to create currently, but can still test what impact their release will have on a smaller community. I plan to continue to study the ethical implications of novel AI techniques, such as graph neural networks that can understand the details of someone's social graph, by looking at a small online community. Thus when the models are available to larger social network platforms, both users and developers will understand what mistakes to avoid.

Human Search Algorithms. Predicting what humans will do can be thought of as a simple task of guessing the next action, but we know that humans can look a few steps into the future. In my work, I have solved this with deep learning techniques[1,4], and demonstrated how other prediction paradigms may be applied to this behavioral modeling task. I plan to explore this task more deeply, model *how* humans search through their space of possible actions. This knowledge will both improve computational system's ability to interact with complex systems and allow for AI systems that can better collaborate with humans by understanding the information that the humans will need.

Useful AI Teachers. I have the building blocks of a system for algorithmically making useful suggestions to humans[1,2,4], and I want to continue to build on this work, develop methodologies for measuring humanness and generalizable techniques for building AIs that can understand people's subtleties. I will expand to different type of tasks and run experiments to verify that people can derive benefit from AI teaching systems. My goal is to build systems where instead of asking the AI what was incorrect or needs to be fixed, normal users can ask of it 'how can I improve?' and rely on the model having a better understanding to determine where to help. Success in this task would significantly improve current remote teaching methods, as it would allow students to get the specific attention they need without the teacher's direct involvement.

[1] Reid McIlroy-Young et al. "Aligning Superhuman AI with Human Behavior: Chess as a Model System". In: *KDD* (2020)
[2] Reid McIlroy-Young et al. "Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess". In: *NeurIPS* (2021)
[3] Reid McIlroy-Young et al. "Learning Personalized Models of Human Behavior in Chess". In: *KDD* (2022)
[4] Reid McIlroy-Young et al. "Mimetic Models: Ethical Implications of AI that Acts Like You". In: *AIES* (2022)