Artificial Intelligence in Education as a Rawlsian Massively Multiplayer Game : A Thought Experiment on AI Ethics

Cowley, Benjamin Ultan

Springer
2022-11-06

Cowley , B U , Charles , D , Pfuhl , G & Rusanen , A-M 2022 , Artificial Intelligence in Education as a Rawlsian Massively Multiplayer Game : A Thought Experiment on AI Ethics . in H Niemi , R D Pea & Y Lu (eds) , AI in Learning : Designing the Future . Springer , Cham , pp. 297-316 . https://doi.org/10.1007/978-3-031-09687-7_18

http://hdl.handle.net/10138/351516
https://doi.org/10.1007/978-3-031-09687-7_18

Artificial Intelligence in Education as a Rawlsian Massively Multiplayer Game: A Thought Experiment on AI Ethics

Benjamin Ultan Cowley, Darryl Charles, Gerit Pfuhl, and Anna-Mari Rusanen

Contents

B. U. Cowley (✉)
Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland

Cognitive Science, Faculty of Arts, University of Helsinki, Helsinki, Finland
e-mail: ben.cowley@helsinki.fi

D. Charles
School of Computing, Engineering & Intelligent Systems, University of Ulster, Ulster, UK
e-mail: dk.charles@ulster.ac.uk

G. Pfuhl
Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway

Department of Psychology, Norwegian University of Science and Technology, Norway
e-mail: gerit.pfuhl@uit.no; gerit.pfuhl@ntnu.no

A.-M. Rusanen
Cognitive Science, Faculty of Arts, University of Helsinki, Helsinki, Finland
e-mail: anna-mari.rusanen@helsinki.fi

H. Niemi et al. (eds.), *AI in Learning: Designing the Future*,
https://doi.org/10.1007/978-3-031-09687-7_18

1 Introduction

Interest in the use of artificial intelligence (AI) within education (AIEd) has grown steadily over the past thirty years, and AI systems are already widely used in a non-teaching manner, e.g. for analytics, monitoring attainment or class planning. Indeed, the OECD firmly advocates the use of AI to measure and improve learning (Kuhl et al. 2019), leading towards digital, data-led governance and AI-based policy-making (Berendt et al. 2017). With the increasing power and ubiquity of computer technology and deep learning algorithms, AI now has the capacity to radically change education. For example, the natural language algorithm GPT-3 (Generative Pre-trained Transformer 3)[1] can help people write code, create websites or apps, co-author stories, summarise legal text and create virtual avatars that chat believably with a person. However, AI-led personalised teaching is in its infancy and carries challenges that must be foreseen and regulated. As yet, there is no methodology to help manage the trade-off between AIEd's possible benefits and challenges (Berendt et al. 2020).

A fundamental problem for AI is the challenge of evaluating AI algorithms in a fair and useful way (Hernández-Orallo 2017b), termed 'explainable AI' (XAI). In the education domain, this problem *prefigures* ethical issues because AIEd implies *also* evaluating the humans interacting with AI, who would otherwise (*sans* AI) be operating according to traditional norms. In other words, XAI in AIEd requires evaluating AI algorithms in a context where the performance is traditionally socially constructed (Latour 2005) and done by humans, for humans. Addressing this problem helps to make AIEd more transparent, which can help tackle the ethical quandaries of AIEd, such as distributive fairness (as defined normatively by, e.g. Rawls 1985). Reliable ways to tackle these issues are important for policy makers and other stakeholders involved in curricula development.

In this chapter, we describe a design for a formal setting, where the general concept of AI-enhanced education is simulated as a massively multiplayer online game (MMOG). The aim is to examine the representativeness of given algorithms for classes of individuals and thereby improve AI transparency, independently of which algorithm is examined. Simulations and games have an extensive track record for teaching and learning within the higher education sector (Lean et al. 2021). In response to the inherent problem of satisfying XAI within AIEd, the MMOG simulation provides a way to make benefit–risk comparisons in multi-stakeholder scenarios, including one which we illustrate explicitly as a thought experiment: the Rawlsian justice game (Rawls 1985) applied to the ethics of AI fairness. Rawls' theory and thought experiment game have been important in political philosophy for many decades, and this chapter is an initial attempt by us to integrate it into research on videogame-based learning.

[1] GPT-3 is a deep learning language model, see https://openai.com/blog/gpt-3-apps/.

In the rest of the chapter, we first describe the theoretical background in Sect. 2, and then Sect. 3 illustrates how the MMOG simulation is designed. Section 4 shows how the simulation integrates a Rawlsian justice game, and we discuss implications and future directions in Sect. 5.

2 Theoretical Background

Teaching is a dynamic and socially interactive process between at least two individuals (Powell and Kalina 2009) and requires adaptation to novelty, uncertainty and change to ensure efficient learning. AI, we argue, can assist human-guided teaching but requires some scaffolding to do so, and the scope of this requirement ranges from the pragmatic (e.g. XAI) to the epistemic.

2.1 Explaining and Evaluating AI

Hernández-Orallo (2017a) describes the crux of the AI evaluation problem: if AI research is the science of making intelligent machines, then algorithms should be evaluated on their intelligence; however, if AI is pragmatically about making machines that perform tasks that would require intelligence if done by humans, their evaluation should be a test of task performance. Thus, the form of the evaluation follows from the scope of the AI: general-purpose AI needs ability-focused evaluation (meaning *cognitive* abilities) and specialised AI needs task-focused evaluation (Hernández-Orallo 2017a).

Most work has been done on task-focused evaluation of specialised AI. Much of this work has had little regard for best practices of human psychometrics (e.g. comparing AI performance to a human reference from a single person) (Cowley et al. 2022). On the other hand, in visual object recognition, for example, (Rajalingham et al. 2018), the best studies are massive and systematic and illustrate great recent progress, as the algorithms become unsupervised and even begin to display biological plausibility (Zhuang et al. 2021). Such work also illustrates one popular method by which algorithms can be judged trustworthy: by human benchmarking. The general approach of benchmarking is central to AI development but has been criticised on grounds that treating a data benchmark as "independent of context, scope and specificity is. . . a false premise for machine learning evaluation" (Raji et al. 2021).

By contrast, human performance benchmarks are implicitly bound to context. For example, in the specialised AI domain of language models, recent work (Lin et al. 2021) reported a human benchmark designed to show model truthfulness (testing the well-known GPT-3 and variants). Results showed that the largest models made most errors, by learning popular misconceptions from the training data—in other words, the most 'powerful' AI was also most prone to learn errors hidden

in the data. Another study (Mohseni et al. 2021) designed a visual recognition benchmark from aggregate human attention data, surpassing benchmarks built on either ground truth image segmentation or human subjective ratings. *These task-focused evaluations illustrate a key issue in AIEd: effective evaluation correlates with ethical evaluation, as both require representative, unbiased, human-grounded training data and/or benchmarks.*

The primacy of task-focused evaluation derives in part from how AI systems typically overspecialise to the task, exemplified by Marcus' (Marcus 2018) list of 10 limitations of so-called 'deep' machine learning[2]: (1) data hungry, (2) limited transfer, (3) lack of hierarchical structure, (4) poor at open-ended inference, (5) not transparent, (6) not well-integrated to prior knowledge, (7) no causal representation, (8) presumes stability, (9) easily fooled and (10) hard to use for engineering. Any or all of these create serious problems in the domain of AIEd. Of course, other families of algorithms exist, but these also often leverage deep learning in some way, and come with their own challenges for evaluation (Henderson et al. 2019).

Even when task-focused evaluation can be done, there is still the challenge of how to use measured performance in a task to evaluate capability, without error-prone extrapolation. Focusing instead on evaluation of ability is not a silver bullet because abilities are constructs that must be defined, requiring a theoretical framework often derived from behavioural sciences. Bhatnagar et al. (2018) reports some work to map out intelligence in a general manner, and Hernández-Orallo (2017a) proposed a kind of universal psychometrics as a possible future solution. Nevertheless, ability- or intelligence-focused evaluation remains a hard, unsolved problem.

In a constrained context such as education, a hybrid approach might be viable given the wide range of preexisting tasks, and the proliferation of psychometrics or other testing instruments, available there. On the other hand, (as noted above) XAI evaluation requires representative data and benchmarks, and obtaining such presents a particular challenge in the education domain. This domain is replete with contra-indications for, e.g. Marcus' list of deep learning vulnerabilities: learning transfer is required, data is hierarchical, learning ablates stability, etc.

The solution we propose, as a thought experiment, is exactly to constrain the domain by setting AIEd within an MMOG. Within such a *simulation* of the classroom, we can experiment with the potential effects of various AI designs. An MMOG-based simulation is a bounded domain with a well-defined application–programmer interface (API), yet nevertheless supports rich, emergent social interaction of players with varied roles. It also does not need to invent novel XAI solutions to individual AIEd problems: rather, the MMOG provides an operating environment where well-structured data and benchmarks can be obtained directly from the game engine.

AI in games has always been a field leader (Laird and Van Lent 2001; Vinyals et al. 2019), and this application domain can be leveraged to illustrate

[2] Not to be confused with deep human learning in education.

how problems of adaptivity and uncertainty can be dealt with in a well-defined context. For example, adaptive AI in games requires two constraints: to maintain logical consistency of game rules and a coherent 'Magic Circle' that preserves player immersion (Huizinga 1949). Games have also been used in XAI, e.g. the Arcade Learning Environment (Bellemare et al. 2013) and the General Video Game Competition (Perez-Liebana et al. 2016), which both consist of collections of game tasks designed to be solved by a single AI agent, and associated evaluations. These works aim to aggregate multiple task-focused evaluations and thereby measure general ability in some sense. Following this approach, the MMOG simulation we propose would 'wargame' various scenarios of AIEd.

2.2 MMOGs, MOOCs and Game-Based Learning

Here, we give background on the kinds of game we envision in our thought experiment. Already 40 years ago, Malone (1981) suggested video games can simultaneously deliver learning and motivation. Kirriemuir and others suggest digital games make excellent motivational tools that promote learning and engagement (Kirriemuir and McFarlane 2004), because they intrinsically motivate players to progress in the absence of extrinsic rewards (Malone et al. 1987) and thus engage the player to master a challenge that can be difficult, prolonged and complex (Charles 2010).

Game *design* also has a lot to offer to learning design, as Gee (2003) outlined with his taxonomy of learning principles in games, which then inspired our own work on learning designs for MMOGs (Cowley et al. 2011). In more recent times, 'gamification' and 'gamefulness' in learning have become popular topics of applied research. Often the focus of these approaches is using games and theories from cognitive and educational psychology to help support and motivate learning—mirroring the long-established use of games in political philosophy (Rawls 1985).

Game playing can be a very social activity, and some of the most popular recent games are only online, including shooter games like Destiny or Fortnite, real-time strategy games like Dota or StarCraft or roleplay games (MMORPGs) like RuneScape and Final Fantasy XIV. A large part of the appeal of multiplayer games is in the strong social bonds that can be built through co-operation and competition in structured play within an 'unreal' environment, each player taking on a role in a fantasy world.

In the early 2000s, MMOGs were a 'natural laboratory' to study how individuals interact online, and proposed as a tool for digitising education (Cowley et al. 2011; Sourmelis et al. 2017). MMOGs enable two features valued in education: role-taking (expressing 'versions' of oneself in different contexts) and groupwork (important for developing skills transferable to the workplace). Furthermore, a multi-user environment provides a richer context for player choice and a wider psychological basis for behavioural variation than single-player scenarios; for example, explicit

competition and collaboration with others, socialising, philanthropy, disruptive behaviour (e.g. 'griefing', 'trolling', cheating) etc.

The MMOG is a useful conceptual construct, not least because it has been so well studied, and serves well as the design for a thought experiment simulation. MMOGs also have one distinct advantage over the newer forms of social online platforms: being games, they naturally conform better to the characteristics of formal games, i.e. they describe the behaviour of rational agents (rationality here defined by the rules of the game, entered into knowingly by the players, viz the Magic Circle Huizinga 1949). This allows us to reason about the behaviour of players with confidence.

2.3 Role of AI in Education

We consider AIEd as incorporating the traditional roles of learners and teachers within a socially constructed educational milieu (Latour 2005). In other words, we start from the assumption that all roles, for human or AI players, for staff or students, are derived from equivalent fundamentals and obtain their unique character through emergence by social construction. This is in line with Actor-Network Theory (ANT) (Latour 2005), which posits that everything in the social and natural worlds exists in constantly shifting networks of relationships. Rather than a predictive theory, ANT provides an empirical 'form of inquiry', which we follow by exploiting the bounded structure and complete access to activity data of MMOGs, to track 'players' and their interactions.

The roles within the classroom are flexible and mutable. Teacher(s), learner(s), and the social group—e.g. the peer group from the point of view of a given learner—sometimes have more teaching and sometimes more learning motivations. That is, teachers are sometimes in training, and thus also learners. And learners sometimes act as teaching assistants or peer mentors, and are thus also teachers. And this conforms to the socially constructed view, since social constructions are goal oriented. In the general sense, the milieu is not defined by fixed, assigned *roles*, but by shifting relational *goals*.

The future of education must now accommodate another role: AI. How AI-driven roles might perturb the socially constructed equilibrium of the classroom is not known *a priori*: in fact every format of the technology can have a different effect. AI-based learning analytics will play a different role to AI instructional agents or to AI agent-based models of individual learners. How should one anticipate or control the ethical goodness of such unforeseen outcomes?

2.4 Ethics of AIEd

From a wider epistemic point of view, AI and other smart technologies change not only the traditional social or physical environments of learning, but also impact the epistemic distribution of labour in classrooms. The role-taking example described above is one example. Thus, AIEd raises a need to evaluate the norms governing the practices of epistemic communities. For example, when cognitive tasks are delegated to machines, it may impact on assessments of 'trustworthiness'. Trust, or reliance, binds the individual epistemic actors into knowledge communities.

Crucially, in AIEd-based knowledge communities, the individuals need not only extend trust to other individuals but also to instruments and equipment they use. That is, individuals should be able to have reliance that epistemic artefacts—such as computers or data analysis methods—work correctly and generate accurate outcomes.

The opacity of contemporary AI applications threatens this binding of reliance and trust. Many current machine learning systems (such as Deep Neural Networks) are so-called 'black box' systems. By definition, we cannot fully explain how such systems work, and thus we cannot fully rely on them as epistemic instruments. This raises a fundamental and deep challenge for the deployment of these technologies as epistemic instruments in knowledge communities (Lo Piano 2020).

There are also many open questions regarding what constitutes transparency or explainability for classroom technologies and what level of transparency is sufficient for different epistemic actors with various positions and roles. For each actor, the interpretation and requirements of 'transparency' may vary. While for a teacher (responsible public sector actor), transparency may require a sufficient understanding of the reliability of a student assessment system, for a student, transparency may mean a comprehensible justification for the decision being made. Or, transparency required to analyse legal significance of unjust biases in learning analytics may mean a different thing than explainability in computer science terms.

Thus, there is a need to develop how we analyse and assess the nuanced aspects of explainability for different actors in different classroom situations. The AIEd-MMOG we consider herein aims to address this need.

3 Methodology and Analysis

In this section, we define a complete schema of an AIEd-MMOG, which we use in Sect. 4 to examine the potential ethical problems of AIEd fairness.

First, we define the setting and the population. The thought experiment proposes launching the AIEd-MMOG in teacher training courses sited within several third-level institutions, wherein prospective teachers learn about the uses and challenges of AI in their future career. This setting provides the following features:

– Players are adults only, avoiding issues related to both developmental mutability and legal/ethical issues of child protection.
– This setting maximises the versatility of role-taking, since trainee teachers can meaningfully embody both teachers and students and consider perspectives on a range of subject-matter disciplines.
– The social power hierarchy is close to 'flat', which helps to prevent unwitting exploitation and avoid undesired influences of power imbalance

3.1 AIEd-MMOG Schematic Technical Definition

The AIEd-MMOG will take the form of an open-world 'sandbox' style game, wherein various tools and toys pre-exist within a single large environment (the sandbox), which allows players freedom to engage as they prefer. This is a similar format as some of the most popular games of recent years, including Fortnite and Grand Theft Auto V. In such settings, avatar API can be run by humans or agent AI—i.e. the actors in the game (avatars) are like robots whose actions are 'programmed' by either human or AI.

The AIEd-MMOG will leverage off-the-shelf technology (i.e. pre-existing and ready to use), such as the Unity game engine, which provides a vast array of software libraries to exploit. This technology will be used to build an environment to support a variety of different learning goals, by packaging learning content as 'mini-games'. Such mini-games can indeed be simple games, or teaching/training tools, or aptitude tests, or hybrids of any/all of the above. Gamified cognitive tests illustrate one way to make such hybrids (Lumsden et al. 2016).

This design ethos of a social online world with embedded modular content has been trialled and evaluated in Cowley et al. (2011) and Cowley and Bateman (2017). Figure 1 shows example screens and architecture from an AIEd-MMOG previously designed by the first author: this example game illustrates how such a game could be structured. Other educational games have also exemplified this design ethos, for example, 'Real Lives: you are the world' (Educational Simulations 2010).

The MMOG content will be versatile due to its modular design, permitting 'minigame' activities to also present moral dilemmas, such as those used to study AI ethics in Sundvall et al. (2021). Compared to such survey-based research, this setting offers the advantage that the dilemmas are lived and not just self-reported on—in other words, participants will not just view a moral dilemma vignette but will face the dilemma on their own behalf.

3.2 Player Models

Within the sandbox-and-minigames environment of our AIEd-MMOG, player behaviour will conform (with some margin of error) to certain predictable patterns

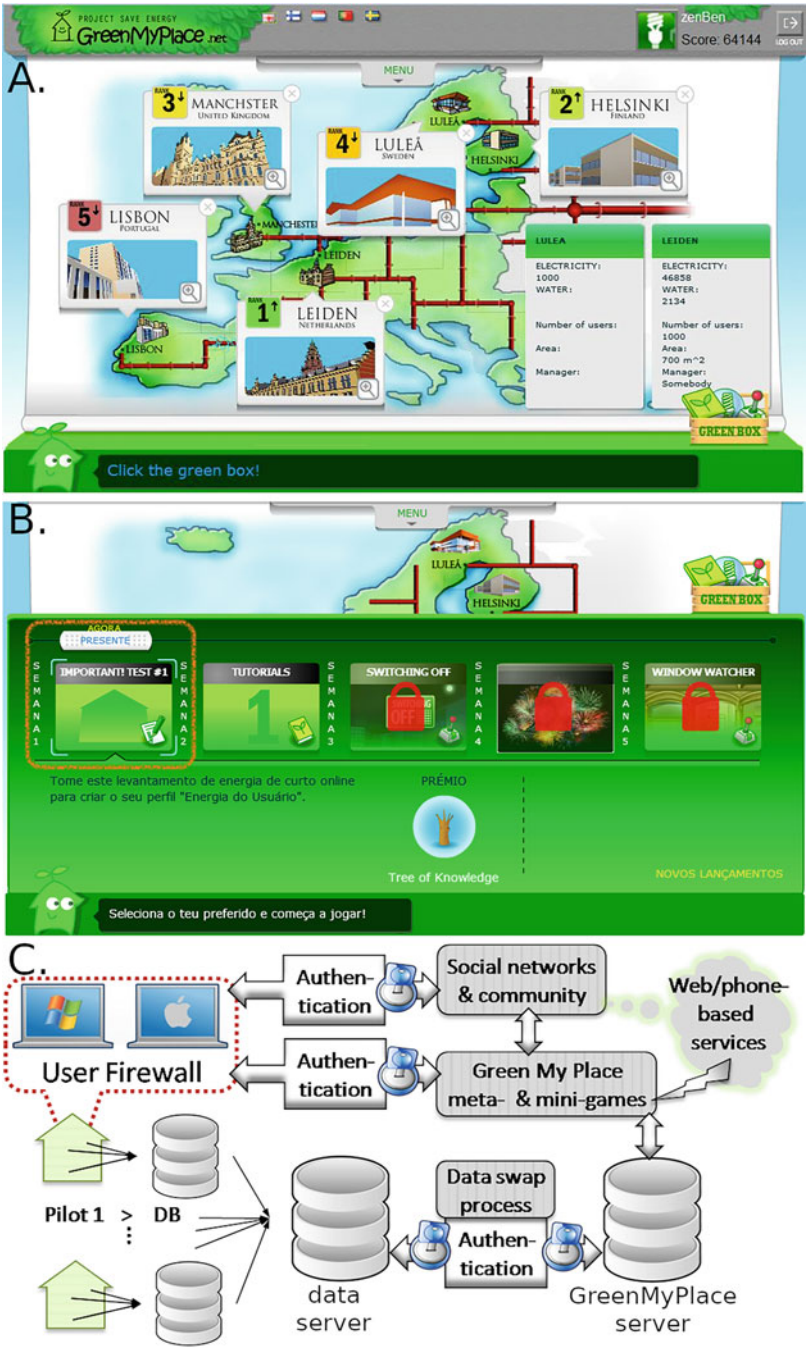

Fig. 1 An exemplar MMOG taken from the first author's earlier work. GreenMyPlace was a massively multiplayer social online game designed to teach concepts of energy efficiency and

because play must conform to how each game was designed to be played, i.e. to game design patterns. This does not mean all players must take exactly the same actions, merely that actions are similar and follow some clustering. This predictability will be exploited to model the types of play behaviour, which can give insight into the player themselves, when tracked over time as a player model. Importantly, insights derived from human players can also be applied to AI agent-based players, since all players use the same API to interact with the game and each other. This helps address the fundamental XAI challenge of equitable evaluation of human–AI activity.

In the MMOG simulation, the abilities that players use to interact are constructed from a hierarchy of tasks. This concept of a hierarchy of tasks that encapsulates the mechanics of a game has been termed *skill atoms* (by Daniel Cook[3]). A skill atom consists of a game action, which results in the application of game rules to change game state in the simulation, and the provision of feedback to the player. Based on this, a process occurs in which the player updates their mental model of the game as a system. The formalism of skill atoms is analogous to a finite-state machine. Furthermore, composition of skill atoms into chains of actions can be used to capture player behaviour (see Fig. 2).

Previously, we showed how such 'skill-atom chains' of behaviour can be linked to player temperament to derive micro-models of play preference, called Behavlets (Cowley and Charles 2016). The Behavlets method leverages domain-expert knowledge of game design patterns, to encode short activity sequences that represent an aspect of playing style or player personality traits (e.g. aggressive or cautious play), which can be mapped to temperament theory. Behavlets have been used to profile players by their play preference (Cowley et al. 2013).

Behavlets can be further analysed as temporally extended sequences called 'B-chains' (Charles and Cowley 2020). The skill-atom>Behavlets>B-chain stack of methods can be considered a hierarchically arranged model of a 'player', each layer trading detail for generality, which when combined serves several purposes:

1. **Efficiency**: Behavlets reduce the dimensionality of game-play data, enhancing algorithmic efficiency and allowing comparison between players in terms of meaningful action

Fig. 1 (continued) promote relevant behaviour change. Panel **a**: the top-level game involved 5 participating pilot locations around Europe which formed the game's 'teams'. Panel **b**: each individual player participated by selecting varied activities from a 'Green Box'. Panel **c**: the game architecture illustrates a model game design for our purposes, providing a controlled, authenticated flow of data from local sites (e.g. classrooms) to online servers, to end-user devices—and back again

[3] https://www.gamedeveloper.com/design/the-chemistry-of-game-design.

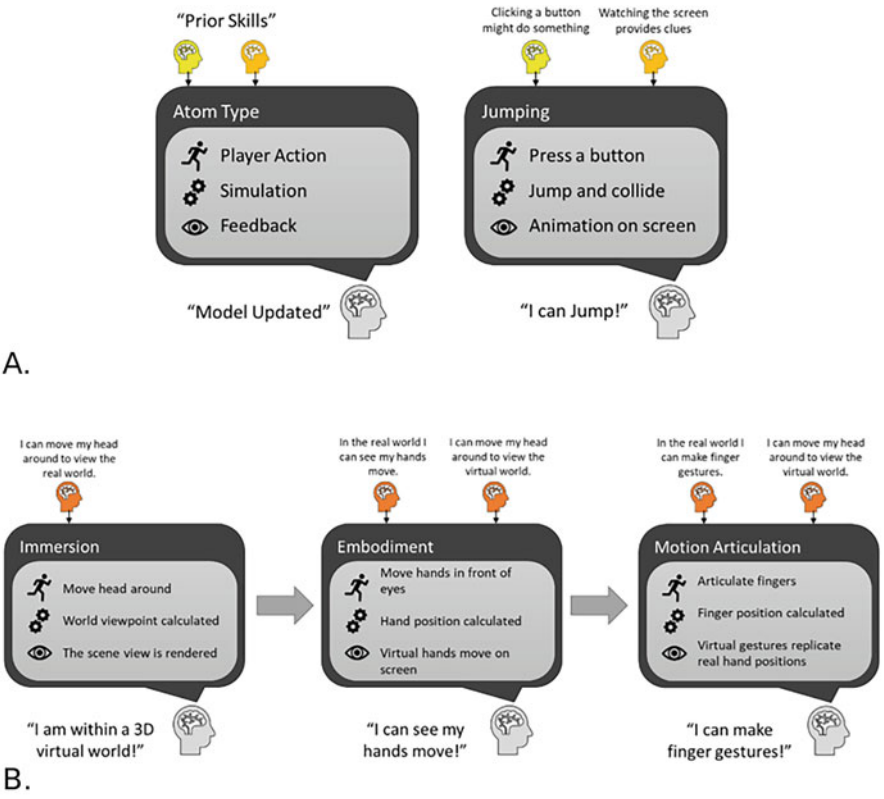

Fig. 2 The concept of a skill atom and their composition to produce skills. Panel **a**: skill atom prototype (left) and jump atom example (right). Panel **b**: an atom skill chain for a natural hand movement game controller in virtual reality

2. **Privacy**: such dimensional reduction means that Behavlets obscure the exact behaviours of individuals, such as key press or mouse movement logs, which have been shown to allow automated identification
3. **Profiling**: clustering human player data permits identification of playing styles (Cowley et al. 2013), which gives insight into AI agent players behaviour

Individual players may have different preferences for their style of play and thus vary in their motivations. We can capture such preferences by the above approach (Cowley et al. 2013); then, using a semi-supervised learning approach based on tracking which Behavlets are triggered by players (human or AI), information is gained on which play preferences are expressed. Similarly, the AIEd-MMOG simulation provides representative data on human attention from the Behavlet and B-chain model layers, which can be used for task-focused benchmarking (within whatever mini-game task the data is taken from).

In Summary, the MMOG simulation for AIEd provides *a terrarium-society environment, wherein interactions can develop according to natural social patterns but bounded by constraints that ensure safety, explainability, reproducibility and transparency of outcomes* (Lo Piano 2020).

4 Findings

In a real educational policy situation, how can the AIEd-MMOG help authorities to decide whether a school or an educational system should deploy a given AI algorithm? Let us consider a concrete example. Suppose you are the Chief Digital Officer in a school district. You are asked to consider whether the region's educational organisation should move from a 'reactive' student guidance system to a 'preventive' guidance system. It would be a novel, sophisticated machine learning system that would help authorised school personnel, such as social workers, to forecast the possible social, cognitive or psychological learning problems of elementary school students. These methods would produce predictions by combining and analysing various sources of student data, including their learning results and, say, medical records. By analysing a large amount of criteria data, high-risk individuals could be identified and prioritised. These high-risk individuals could proactively be invited to meet with school tutors, social workers, counsellors or psychologists, to get guidance and help.

Obviously, the preventive system would have many positive possibilities, including potential to improve overall well-being of students. Furthermore, it might allow better student supervision, supportive actions and impact estimation. At the same time, the preventive system raises several legal and ethical issues regarding privacy, security and use of data. It raises the fundamental question of justification: do the authorities have a principled right to use private and sensitive data for identifying high-risk students, and if so, to what extent? And, if these systems are used, will individuals be treated in an equal way? What exactly *is* equality in this context, where different individuals have different needs and roles? How to distribute whatever resources are inherent in deployment of an AI algorithm, to ensure a fair and just outcome for all students, when the algorithms and their deployment mechanisms are not (cannot be) transparent?

AIEd-MMOGs could provide a formal setting for simulating these situations, where individuals do not start from the same position, and there are individual differences that matter. These simulations can be used to bring together, e.g. philosophical ideas on distributive justice, with an active, instantiated environment that facilitates testing of various alternative approaches to real-world scenarios.

4.1 Rawlsian Justice Game

According to John Rawls' theory of justice, the distribution of resources should maximise the benefits to the members who start with minimal resources (Rawls 1985). The most important principle of fairness is to ensure that the 'least advantaged' members of society will benefit and not be harmed. The distribution of resources that maximises the benefit to the members who start with minimal resources is the *maximin distribution*. Rawls' idea was that individuals in a society must choose their preferred distribution function with no foreknowledge of their own status in the society: a feature dubbed the *veil of ignorance*. From behind the veil of ignorance, Rawls claims, individuals will tend to select the maximin distribution.

Howe and Roemer (1981), among others, have described how Rawlsian justice can be modelled as a game, in their case for economic distribution. Such *Rawlsian justice games* (RJGs) have been used as classroom teaching tools in a variety of disciplines including political science and economics (Alden 2005), where students debate and select various distributive principles.

In the Howe and Roemer (1981) model, individuals from a population $p \in P$ will each receive an *endowment* α (which ranges in $[a, b]$), under some probability distribution $f(P)$. Now, the veil of ignorance prevents any p from knowing about $\alpha(p)$, but they may know about f. Endowment is converted to income Y (aka "the good") via some production function. Redistribution of incomes, which is the scheme that the population may choose via the game, is modelled as a tax τ.

Howe and Roemer (1981) then describe the *incentive problem*: as $\tau(p)$ rises, p will produce less *pre-tax* income. This is modelled by the production-incentive function $g(\alpha, \tau)$, which maps endowment and tax to income produced, which is $(1-\tau)g(\alpha, \tau)$. Behind the veil of ignorance, the population know g exists, but not how it operates. Howe and Roemer (1981) go on to define tax schemes and the maximin distribution within this model, the details of which are not critical here. Then they describe how a game can be structured: behind the veil of ignorance, every p will aim to choose a tax scheme $\tau(p)$, such that, after endowments α are assigned and post-tax incomes Y realised, no *coalition* of players $p_{i..j} \subset P$ can improve on Y by attempting to draw again from f. Multiple draws on f may be expected before an equilibrium is reached. In other words, a key aspect of the RJG is that it will "*... allow an individual or coalition of individuals to express its dissatisfaction with a particular income distribution by hypothetically withdrawing from society, and testing whether under the rules of the game it can improve the lot of its members*" (Howe and Roemer 1981).

We suggest that such a permutation testing mechanism can function as an evaluation of AI algorithms. To implement that we envision a selection of social learning mini-games within the MMOG, each game distinguished by variants of an AI algorithm. This social learning mini-game can be anything, so long as it prescribes some type of multi-user engagement (such as group-wise problem-based learning, PBL) and supports standard testing of learning outcomes (to enable

quantification and then automation of evaluation). Also critical is that in the AIEd-MMOG design described, performance in mini-games contributes to overall performance of one's 'team' (physical institution), thus the immediate outcome is important on the macro scale.

4.2 AIEd-MMOG Rawlsian Justice Game

To adapt the mechanics of the RJG to the AIEd-MMOG environment and obtain an *AIEd-RJG*, we must further answer the question: what is justice in this AIEd domain? What does the social contract govern, and/or what is justice regulating (since it is not monetary income)? To answer this, the implementation should map from the traditional concepts of the game to concepts that make sense in the domain. For 'income' we map from income as money to income as learning (measured by standardised test). For endowment, we map from endowment-as-social-status to endowment-as-*representation*, i.e. how representative was the training data for each person? In terms of our concrete example of preventive guidance, this will translate to accurately can a student be assessed based on the representation of their characteristics in the data.

Thus, justice will be defined as relevance of the AI to each person's given background and ability to learn and test well. If the AI does well for a given student, then the student should elect to continue with that particular algorithm; or conversely, they may elect to switch to another environment with an alternative algorithm. *However*, because mini-games within the RGJ depend on group-based PBL, then switching to another environment can only maximise learning outcomes if many other players *also* switch.[4] In practice this will tend to favour joint action by 'coalitions', as in the original RJG (1981).

Thus, our proposed AIEd-RGJ will function at the 'level' of the MMOG, accumulating data on the 'goodness' of separate draws on f via the mechanism of players' choice of mini-games. The adapted AIEd-RJG will then operate as follows:

- Individuals $p \in P$ become learning agents $l \in L$.
- Endowment α becomes representation $\bar{\alpha}$.
- Income Y becomes learning (test score) $\bar{Y}$.
- Endowment probability distribution f becomes the probability of occurrence of L in the AI training data, denoted $\bar{f}$.
- The tax rate τ becomes the cost of mentoring peers in $l_{i\ldots j} \subset L$, denoted $\bar{\tau}$.
- Production-incentive function g becomes the personal learning incentive $\bar{g}$[5]
- The personal learning outcome is then $(1 - \bar{\alpha})g(\bar{\alpha}, \bar{\tau})$

[4] This mechanic is similar to when players in commercial MMOGs switch between servers.

[5] The interpretation of $\bar{g}$ is that intrinsic motivation to learn is weighed against the need to help peers to lift up team performance.

$\bar{f}$ and $\bar{g}$ are both unknown in AIEd-RJG because of *AI non-transparency*, corresponding to the veil of ignorance! They can be estimated from sufficient repeated plays (draws on $\bar{f}$). In other words, by actually playing the game, L creates data to estimate $\bar{f}, \bar{g}$. Play will end when no one wishes to try for better learning scores by sampling again (hoping for better representation).

However, the volume of play which is sufficient might be onerous for human players, especially recruited from a teacher training programme. To help solve this, the human-based play data can be supplemented with AI agent-based play, by training AI agent algorithms to play in a manner that emulates human playing style based on the 'seed' games played initially by humans. The techniques to do this are beyond the scope of this paper, however they rely on the methods described in Sect. 3.2 for modelling player personality, i.e. skill atoms, Behavlets and B-chains.

4.3 AIEd-RJG for AI Evaluation

Returning to the question of XAI evaluation, our AIEd-MMOG simulation follows from prior game-based AI evaluation work (Bellemare et al. 2013; Perez-Liebana et al. 2016), suggesting that ability can be evaluated from the aggregate of task evaluations. Our simulation adds the capability to assess (a) social influences on task performance from player-to-player interactions and (b) the representativeness of given algorithms for classes of individuals. This all aims to improve AI transparency, independently of which algorithm is used: although we cannot always see inside the black box, we *can* forecast how it behaves.

Note that in this approach, two kinds of AIEd algorithm can actually be tested: (a) agent-based AI that plays the game alongside humans or (b) the analytics/oversight algorithm that models players and distributes the 'social good' (thus conforming to the concept of a market principle in original RJG).

5 Discussion/Synthesis

In this chapter, we have presented a thought experiment on how to use an MMOG simulation to study AIEd deployment solutions, focusing on the fundamental challenge of explainable AI, examined through the lens of Rawlsian distributive justice.

As stated by Schulzke (2012), 'by taking a concept like distributive justice out of the realm of theoretical speculation and making it part of a simulation, games provide an excellent means of recontextualising the problem by giving players firsthand, concrete experience of that problem'.

Schulzke in fact examined the educational game *Real Lives* from the perspective of an RJG, thus linking it to our thought experiment by design format. That work focused on natural justice, not AIEd, and within *Real Lives* the Rawlsian lesson is never explicit. However, Schulzke's commentary shows the relevance of an MMOG-format game for examining Rawlsian concepts. Rawlsian justice has also been modelled in the context of AI ethics (Leben 2017), although (to our knowledge) our work is first to situate an RJG within AIEd.

5.1 Implications

Responsible AI requires that choices and decisions be explicitly reported and open to inspection, i.e. they meet the ART principles: Accountability, Responsibility and Transparency (Dignum 2021, p-3).

Accountability includes that all stakeholders are involved in defining the moral values and societal norms that AI represents (is designed for). *Responsibility* encompasses the user's relation to AI, already at development and also when using the system. *Transparency* refers to describing, inspecting and reproducing how the AI system learns to make decisions and adapt to its environment, thus ensuring trust. Transparency also refers to explicitly and openly describing data sources for training, development processes and stakeholders. Not meeting ART requirements can lead to stakeholder dissatisfaction and 'bandaid' fixes, such as post hoc regulation.

The AIEd-MMOG meets all ART principles. Accountability, because the environment combines top-down designed constrains on actions with a bottom-up process of social construction to shape the games' moral norms. Responsibility, because building the human–AI relationships on a foundation of well-defined XAI permits comprehensive comparable evaluation. Transparency, because the MMOG is a strictly bounded environment where code is open, data has clear provenance, and actions cannot be hidden—they are even associated with action-motivations through the Behavlets and with action-context through the B-chains.

What is more, the setting provides the opportunity to explicitly represent varied moral stances as minigames, which allow human or AI players to demonstrate their own values as choices.

Finally, given our aim of supporting XAI for more transparent, interpretable and ethical AIEd, note that the MMOG simulation facilitates *reproducible* AI (Pineau et al. 2020), compared to deployment in a live classroom.

5.2 Future Outlook

Successful teaching relies on pedagogic rights and teacher–student relationships governed by enhancement, participation and inclusion (Reiss 2021). *Enhancement*

is education for critical thinking. *Participation* means that the users have the right to be separate and autonomous and not subsumed with the system. *Inclusion* facilitates representative democratic structures, i.e. avoiding dominance of commercial or governmental providers. These pedagogic rights align with acting morally in the humanist sense (2021). A corollary is that any AI system should be made for, but also by, the users who then decide which AI systems are used, and how.

> Conscious and well-informed... individuals will create a solid foundation for responsible and positive uses of AI systems and digital technologies more generally, and strengthen their personal skills on cognitive, social and cultural levels. This will not only increase the available talent pool, but also foster the relevance and quality of research and innovation of AI systems for society as a whole.
>
> (European Commission, n.d.)

These rights correspond to large-scale implementation demands, touching on the AIEd challenges discussed above. By facilitating some small progress towards tackling those challenges, our AIEd-MMOG would allow potential issues to be identified and without running expensive and time-consuming live trials.

6 Conclusions

Although human-performed evaluation in education is sometimes imperfect, it is also important to consider that AI evaluation can be biased, leading to problems of underestimating AI systems or setting too high a bar on them (Buckner 2021). We have described a thought experiment aimed at addressing this dual evaluation issue within the new frontier of AIEd. The proposed AIEd-MMOG is simply a constrained and well-defined *setting* for AI to enter education, a proposed set of features that facilitate bottom-up/task-focused XAI evaluation within a social milieu with deployed AI. In this setting, we have shown how an RJG design could improve AI transparency by estimating how representative is a given algorithm for various classes of individuals.

References

Alden, L. (2005). Econoclass: The distributive justice game. Retrieved 2 December 2021, from http://www.econoclass.com/distjusticegame.html

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research, 47*, 253–279.

Berendt, B., Littlejohn, A., & Blakemore, M. (2020). AI in education: Learner choice and fundamental rights. *Learning, Media and Technology, 45* (3), 312–324.

Berendt, B., Mitros, P., Shacklock, X., Blakemore, M., Littlejohn, A., & Kern, P. (2017). *Big data for monitoring educational systems*. Publications Office of the European Union.

Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B. S., Ó hÉigeartaigh, S., Martínez-Plumed, F., Price, H., Shevlin, H., Weller, A., Winfield, A., & Hernáandez-Orallo, J. (2018). Mapping Intelligence: Requirements and Possibilities. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017* (pp. 117–135). Springer International Publishing.

Buckner, C. J. (2021). Black Boxes, or Unflattering Mirrors? Comparative Bias in the Science of Machine Behavior. *The British Journal for the Philosophy of Science*.

Charles, D., & Cowley, B. U. (2020). Behavlet Analytics for Player Profiling and Churn Prediction. In C. Stephanidis, D. Harris, W.-C. Li, D. D. Schmorrow, C. M. Fidopiastis, P. Zaphiris, A. Ioannou, X. Fang, R. A. Sottilare, & J. Schwarz (Eds.), *HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games* (pp. 631–643). Springer International Publishing.

Charles, T. (2010). *Enhanced e-learning engagement using game absorption techniques ELEGANT* (Doctoral dissertation). University of Ulster.

Cowley, B., Charles, D., Black, M., & Hickey, R. (2013). Real-time rule-based classification of player types in computer games. *User Modeling and User-Adapted Interaction, 23* (5), 489–526.

Cowley, B., Moutinho, J. L., Bateman, C., & Oliveira, A. (2011). Learning principles and interaction design for 'Green My Place': A massively multiplayer serious game. *Entertainment Computing, 2* (2), 103–113.

Cowley, B., & Bateman, C. (2017). Green My Place: Evaluation of a Serious Social Online Game Designed to Promote Energy Effcient Behaviour Change. *International Journal of Serious Games, 4* (4), 71–90.

Cowley, B., & Charles, D. (2016). Behavlets: A Method for Practical Player Modelling using Psychology-Based Player Traits and Domain Specific Features. *User Modeling and User-Adapted Interaction, 26* (2), 257–306.

Cowley, H. P., Natter, M., Gray-Roncal, K., Rhodes, R. E., Johnson, E. C., Drenkow, N., Shead, T. M., Chance, F. S., Wester, B., & Gray-Roncal, W. (2022). A Framework for Rigorous Evaluation of Human Performance in Human and Machine Learning Comparison Studies. *Scientific Reports, in press*.

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education*.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. Palgrave Macmillan.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2019). Deep Reinforcement Learning that Matters. *arXiv:1709.06560*.

Hernáandez-Orallo, J. (2017a). Evaluation in artificial intelligence: From taskoriented to ability-oriented measurement. *Artificial Intelligence Review, 48* (3), 397–447.

Hernández-Orallo, J. (2017b). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.

Howe, R. E., & Roemer, J. E. (1981). Rawlsian Justice as the Core of a Game. *The American Economic Review, 71* (5), 880–895.

Huizinga, J. (1949). *Homo Ludens: A study of the play-element of culture*. Routledge.

Kirriemuir, J., & McFarlane, A. (2004). Report 8: Literature review in games and learning. *Futerelab Series*.

Kuhl, P. K., Lim, S.-S., Guerriero, S., & Damme, D. v. (2019). *Developing Minds in the Digital Age*.

Laird, J. E., & Van Lent, M. (2001). Human level AI's killer application: Interactive computer games. *AI Magazine, 22* (2), 15–25.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Lean, J., Moizer, J., Derham, C., Strachan, L., & Bhuiyan, Z. (2021). RealWorld Learning: Simulation and Gaming. In D. A. Morley & M. G. Jamil (Eds.), *Applied Pedagogies for Higher Education: Real World Learning and Innovation across the Curriculum* (pp. 187–214). Springer International Publishing.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19* (2), 107–115.

Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958*.

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications, 7* (1), 1–7.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games, 4* (2), e5888.

Malone, T. W., Lepper, M. R., Snow, R., & Farr, M. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning and instruction: III. conative and affective process analyses* (pp. 223–253). Hilsdale.

Malone, T. W. (1981). What makes computer games fun? [for education]. *BYTE, 6* (12), 258–277.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv:1801.00631*.

Mohseni, S., Block, J. E., & Ragan, E. (2021). Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. *26th International Conference on Intelligent User Interfaces*, 22–31.

Perez-Liebana, D., Samothrakis, S., Togelius, J., Schaul, T., Lucas, S. M., Couëtoux, A., Lee, J., Lim, C.-U., & Thompson, T. (2016). The 2014 General Video Game Playing Competition. *IEEE Transactions on Computational Intelligence and AI in Games, 8* (3), 229–243.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv:2003.12206*.

Powell, K. C., & Kalina, C. J. (2009). Cognitive and Social Constructivism: Developing Tools for an Effective Classroom. Education, *130* (2), 241–250.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience, 38* (33), 7255–7269.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (2021). AI and the Everything in the Whole Wide World Benchmark.

Rawls, J. (1985). Justice as Fairness: Political not Metaphysical. *Philosophy & Public Affairs, 14* (3), 223–251.

Reiss, M. J. (2021). The use of AI in education: Practicalities and ethical considerations. *London Review of Education*.

Schulzke, M. (2012). Using Video Games to Think About Distributive Justice. *The Journal of Interactive Technology and Pedagogy*, (2).

Sourmelis, T., Ioannou, A., & Zaphiris, P. (2017). Massively Multiplayer Online Role Playing Games (MMORPGs) and the 21st century skills: A comprehensive research review from 2010 to 2016. *Computers in Human Behavior*, 67, 41–48.

Sundvall, J., Drosinou, M., Hannikainen, I. R., Elovaara, K., Halonen, J., Herzon, V., Kopecky, R., Kŏsová, M. J., Koverola, M., Kunnari, A., Perander, S., Saikkonen, T., Palomäki, J., & Laakasuo, M. (2021). *Innocence over utilitarianism - Heightened Moral Standards for Robots in Rescue Dilemmas PREPRINT* (tech. rep.). PsyArXiv.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., . . . Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575* (7782), 350–354.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences, 118* (3).