Ethical Considerations for Player Modeling

Benedikte Mikkelsen and **Christoffer Holmgård** and **Julian Togelius**
Game Innovation Lab, Tandon School of Engineering, New York University
benediktemikkelsen@gmail.com, christoffer@holmgard.org, julian@togelius.com

Abstract

In this paper we discuss some of the ethical challenges that may arise from player modeling. Player modeling is used in modern games e.g. to enable various kinds of game play, to optimize games for specific players, and to maximize the monetization of games. In this paper, we propose that applying player modeling implies serious ethical questions, since it impacts how players spend their leisure time and money, affects their social relations, and changes computer games as ethical artifacts. We source categories of ethical issues in the application of artificial intelligence (AI) from work on AI ethics and using these we provide several specific examples of ethical issues in player modeling. Building from the examples, we suggest establishing a framework for understanding ethical issues in player modeling and we propose a number of methodological approaches to address the identified challenges.

Introduction

Player modeling has become an integrated part of digital games in recent years. The concept of player modeling refers to the study of computational models of players in games. This includes modeling both static features of players, e.g. personality, cultural background, gender, and age, and dynamic phenomena that occur when players interact with games, e.g. playing style and in-game choices. Player modeling operationalizes and captures aspects of player behavior, preferences, traits, or all of the above, and model these to describe, classify, or predict them (Yannakakis et al. 2013; Smith et al. 2011). This can be seen as one use case of *game data mining*, where large data sets generated by or within games are analyzed (Drachen et al. 2013).

In game development, player modeling is typically approached through local and telemetric *game analytics* (El-Nasr, Drachen, and Canossa 2013) and the outcomes used in a modeling framework. The output of these models is typically used to decide what the player experiences when during game play: From what parts of the game are seen, to the intensity of the challenge in the game, to what ads are presented, to name a few examples (Yannakakis et al. 2013). The decisions drawn from the models may be made during the design of the game, before it is released, or they may be integrated into the released game as an integrated aspect of the game, continuously making decisions over the course of the lifespan of the game. Either way, when player modeling is applied in the design of a game, certain decisions about the player experience is delegated to the automated models, rather than being directly implemented by human designers.

Outside of games specifically, in the field of applied machine learning and artificial intelligence in general, recent years has seen an increasing interest in understanding the implications when decision-making previously carried out by humans is delegated to and augmented by automated models (Crawford et al. 2016). Bostrom and Yudkowsky (2014) capture the core of this concern in the following quote:

> "But when AI algorithms take on cognitive work with social dimensions—cognitive tasks previously performed by humans—the AI algorithm inherits the social requirements." (p. 916.)

Recent work in the ethics of modeling and AI has shown how large-scale modeling almost always incorporates values, intentional or unintentional, of the modelers through either the data sets sampled or the modeling methods applied (Sweeney 2013; O'Neil 2016). For example, the "Word2Vec" algorithm, which maps words to points within a large vector space based on a large corpus of texts, has been found to capture and reproduce sexist language in the texts it is trained on (Bolukbasi et al. 2016).

These problems are exacerbated by the difficulty of inspecting and interpreting learned models. The black box nature of the automated models has raised concerns that mistakes or biases in the systems may go undiscovered and even be amplified by feedback mechanisms. For example, investigations into algorithms used in the fields of predictive policing illustrate how models trained on biased data aided by feedback loops results in discriminatory decisions (Lum and Isaac 2016).

Meanwhile, work on the ethics of computer games has shown how games are inherently ethical systems that constitute a form of communication between a game designer and a game player, or between multiple players, where the designer defines the rules of a space in which the players may choose how to act (Sicart 2011). Game designers thus take actions with ethical repercussions, and have some ethical responsibility. When part of the design is based on learned

models, some of the responsibility falls on whoever constructed the model or learning algorithm. In other words, the AI researcher and professional designing the player modeling system is ethically responsible.

This paper explores the ethical implications of player modeling and attempts to outline a way ahead for recognizing and addressing ethical issues that may arise.

In the following section, we start out by reviewing work on the general ethical implications of modeling and automated decision making. We then proceed to review work on computer games as ethical objects and spaces. Following this, we visit a number of specific, contemporary applications of modeling and AI in games today. Finally, we will suggest ways to address the ethical challenges for the use of modeling in computer game design and production.

Related Work

A growing body of work discusses the ethical issues connected with modeling and artificial intelligence; we will give an overview of this below. We will also discuss research on the ethics of computer games and on the connection between games and non-game reality.

Ethics of Artificial Intelligence

Bostrom and Yudkowsky provide a non-exhaustive list of socially important properties or virtues that may be considered whenever an AI takes over cognitive work with social dimensions. The list includes *"responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream in frustration"* (Bostrom and Yudkowsky 2014).

Responsibility refers to the idea that the decisions made by AI systems should be attributable to responsible individuals or organizations. There is a risk that responsibility may be averted by pointing to the decisions of models rather than the parties responsible for implementing these models.

Transparency describes the fact that decisions made by AI systems should be open to inspection or at least explainable to the largest extent possible.

Auditability refers to the fact that mechanisms for ensuring that AI systems act as intended are available.

Even if systems cannot be made open to public inspection, it should be possible for trusted professionals to inspect and audit them.

Incorruptibility refers to systems and models being robust against willful manipulation from external parties, by being robust against attack with e.g. malicious input data.

Predictability refers to the outputs of systems being predictable for users so that similar actions or conditions yield similar outputs over time.

Later, we move on to applying these different virtues and risks of AI to problems particular to player modeling and profiling. However, before addressing this, we briefly review a number of identified ethical hazards of modeling, drawn from management science.

Hazards of Models

Problematic aspects of modeling in social contexts have been a point of discussion for several decades, especially within the field of management science. Wallace (1994) identifies three ethical *hazards of models*, listed below. These categories operate closer to the modeling level than the issues identified by Bostrom and Yudkowsky and serve as a useful complement when examining the ethical challenges in the applications of models:

Assumption of routine and normality refers to the fact that models are typically bad at handling exceptional situations or outliers without expert help. While modern high-dimensional and/or non-linear modeling methods, such as e.g. support vector machines, matrix factorization or (deep) neural networks, combined with data subdivision models such as various variants of clustering, have proven useful in creating models much more sensitive to variations in data sets, the issue remains that even modern data-driven models respond to typical patterns in the data sets they are trained on. To the extent that sampled data sets are inadequate representations of reality, which often exceedingly turns out to be the case for social data, individuals who find themselves at the borders of clusters or as outliers to the class that they belong to, are at risk of being misrepresented by models acting upon them. In a recent example, a model of facial beauty was trained on predominantly fair-skinned women, and would therefore not rate dark-skinned people as beautiful (Levin 2016). If you're a a bad fit for the model, the output of the model will most likely be a bad fit for you. This might be particularly problematic for minorities and other marginalized groups. Exacerbating this problem is that the real world is a non-stationary system, where what was exceptional when a model was learned might be the new norm some time later.

Typing/classification is an issue related to the problem of assuming normality. When normal types are extracted from large data sets, individual variation is necessarily removed from the models, stripping individuals of identifying aspects that are unique to them. If models are constructed in the wrong way, it is quite possible that features that are important from the phenomenal perspective of the individual are removed, if they are not considered important a priori from the perspective of the designer or the modeler.

Type III errors. While the definitions of Type I (false positive) and Type II (false negative) errors are well known in science, Wallace (1994) defines a related third category. Type III errors are used to describe when otherwise well-vetted and validated models are applied to data sets stemming from an inappropriate population—simply put the misapplication of good models to inappropriate data. An relatively innocuous example could be using a model for dynamic difficulty adjustment trained on only expert players for novice players. Unless the model transfers perfectly onto lower skill levels, it's not given that the novice players would enjoy or need the same adjustments as the expert players. The model might be well-built, but applied to the wrong audience it would not have the desired effects.

Ethics of Computer Games

Early inquires into the ethics of computer games focused on the effects that computer games might have on players in general, with less focus on ethical nature of game play

itself (Larsson and Dodig-Crnkovic 2004). More recently, Sicart (2011) published a comprehensive investigation into the ethics of computer games and game play. He observes that

> "Computer games are designed experiences in virtual environments with rules and properties that, in general, cannot be adapted or corrected by their users." (p. 15)

and that

> "The task of the developer, then, is to create behaviors in players by means of constraining and encouraging their actions. This task is, almost by definition, an ethical task, and as such game developers have to both be aware of and bear the responsibility for the ethics of computer games as designed objects." (p. 47)

However, not all players are fully aware of how these rules are represented, evaluated and applied—for many players the internal rules of a game are opaque to the point of being a black box, which increases the impact of the rule system:

> "By not showing how the games' rules are enforced, digital games tend to strengthen the supremacy of the rules system in the experience of the game." (p. 27)

Zagal, Björk, and Lewis (2013), in their work on dark game design patterns, show how the rule structures of games might be used in ways that arguably do not align with the interests of the players by forcing or luring players into e.g. undertaking excessive grinding, paying to skip content, paying to win over other opponents, or to expend social capital in order to progress in a game, by inviting friends from social networks. In many of these these cases the rules that implement these dark patterns are relatively static, general, and not personalized nor adaptive. However, once the rules that drive the game play experience are driven by automatic systems and models derived from large data sets, enabling precise and individualized predictions, arguably this supremacy over the player is amplified and an ethical responsibility is held by the modeler as well as the designer.

Building from the related work cited above, it can be argued that the relation between the player and the designers and producers of a game is inherently ethical and strongly structured through the dynamics implemented by the rules of a game. As these dynamics are increasingly shaped and made adaptive by player modeling, this relationship grows more imbalanced: the power of the designer and the modeler grows as they gain more information about the player through telemetry, analytics, and modeling.

What makes these questions more than an academic concern is the impact they have outside the game play of the game itself. If the ethical relationship was constrained to the world of the game and its game play, it might be argued that players who were dissatisfied with the rules of a particular game are free to leave it at any time and that a free market for computer games would ensure that the interests of players are served.

While players are free to leave computer games and stop playing whenever they want, doing so may have significant real-world consequences and costs that reach beyond the world of the game itself, further increasing the ethical responsibility of modelers and designers. In the following section we argue for why we believe this is the case and outline some of the ways in which games impact players outside of the game itself.

Real Life in Digital Games

Historically, games have been thought of as taking place in "Magic Circles" that delineate play space from the rest of life (Huizinga 1955; Salen and Zimmerman 2004). Recently, and especially as games have started permeating more aspects of life while become continuous and networked, the notion of games taking place in separated spaces has been critiqued and abandoned by some in favor of seeing them as activities situated in and related to players' life in general (Woodford 2008). This is supported by a wide range of ethnographic (Pearce, Boellstorff, and Nardi 2011), sociological (Taylor 2009), and economic (Castronova 2001) research into how games provide significant meaning to the individual and provide a platform for contact and relationships between people who might not otherwise be able to meet or interact. Games and digital play spaces provide spaces where players can experiment with identity and self expression at their own accord. They have become important individual and social spaces where players invest significant amounts of money, time, and social engagement.

Games are also increasing playing a role outside the entertainment industry as they are incorporated into work environments for assessment (Holmgård, Togelius, and Henriksen 2016), the health sector as tools for rehabilitation and diagnostics (Kato 2010) and in education for teaching (Gee 2005) and assessment (Shute and Ventura 2013).

Ethical Challenges in Player Modeling

As noted in the sections above, games are multifaceted environments with complex relations between designers, modelers, and one or more players. Models operating on games or the players of games can have consequences both at the individual level, when a decision is made about a single person, or at the aggregate level, when decisions are made that influence groups of players or the total player population.

At the individual level, models and profiles in games may act on players in a multitude of ways that demand ethical consideration. These include collecting personal or private data, optimizing players' spending in a game, banning players from accessing a game, creating profiles of individual players, making claims about training or assessing individuals, or adapting the game to the individual player.

At the group level player models and profiles may have a significant impact on the balancing of games and matching of players, which determines how players are able to meet and interact in ways mediated by computer games.

In this section we identify a number of specific aspects of computer games where player modeling may encounter ethical issues worth of scrutiny. The list of identified issues is in no way exhaustive, but is intended to be a starting point for identifying ethical issues in player modeling and profiling.

Monetization and Churn Prediction

One of the topics of ethical concern that has received the largest amount of discussion in the public, among game developers, and in the academic literature, is perhaps the topic of player monetization and churn prediction (Alha et al. 2014). As free-to-play games (where the main game is free, but players can choose to pay for items or services along the way as they play) have become a more prolific business model in the games industry, so has the effort directed toward modeling when players decide to spend money inside the a game, and whether players will stay engaged with a game (El-Nasr, Drachen, and Canossa 2013). In free-to-play games, the ability to progress through the game, or the speed at which the player progresses, is partially a function of the amount of real-world money spent. This excludes some players of lesser monetary means from participating at the same level as more affluent players.

As the models for predicting player's in-game spending habits may not necessarily include any information on the player's financial situation, this could possibly lead to predatory loops where models are responsible for keeping players in a situation that is not in their best interest. Players who have already invested time and money in a particular game may be more susceptible to the sunk cost fallacy and may seek to protect their investment by responding to interventions guiding them toward spending and playing.

Because of the potential economic benefits for the developers, they are naturally motivated to avoid the kind of the hazards Wallace (1994) warn about when modeling for monetization optimization and churn prevention. However, the models are typically kept as trade secrets and are not made available to players. To the extent that they are published, e.g. at academic venues, these models are only accessible to a selected crowd with a high technical literacy. So looking at these processes through the framework of virtues quoted by Bostrom and Yudkowsky, the models are typically opaque to the players affected by them, hard to audit, even by public instances, and unpredictable to the player subjected to them—unless the player reverse-engineers the model through experience. (Online wikis and message boards devoted to particular games sometimes feature attempts at second-guessing player models.)

This means that the ethical responsibility of monetization and churn prevention models could be improved by making information about the operations of these models available to the player base. This would allow players to make an informed choice before engaging with games applying such models. Unfortunately, this may run counter to the market-based incentives of companies offering free-to-play games and as such may be considered unlikely to happen without external incentives such as regulation.

Banning

Another specific example where models and profiles exert a large amount of power over players is the case of flagging and banning players from games. Modern multiplayer games often have policies in place for preventing harassment or generally preventing players from exhibiting behavior that has been deemed unacceptable in that particular game. This could include behaviors such as using profanity against other players or using local assistive software to play a game, such as e.g. *aim-bots*. For successful games, the number of players and actions that must be evaluated often exceed what is feasible by manual means, so modeling and profiling is applied to implement standards automatically. This puts a model in the place of a human arbiter for the purpose of evaluating whether players are acting in accordance with the policy in place.

The ethical responsibility incurred by implementing a banning model can be considered double-edged. On the one hand the designers and modelers have a responsibility for protecting their player base against toxic influences from players who might harass or embarrass other players. On the other hand, especially multiplayer games may often provide important social contact to players for whom parts of their social network is only accessible through the game. Additionally, players may have invested hundreds or thousands of hours into a game, possibly along with real-world money, and banning them from accessing the game world effectively nullifies the value of this investment. This underscores that any automatic banning system should be constructed with ethical consideration as the system exerts strong power over individual players in order to protect the perceived interests of many other players. Viewed from the perspective of Bostrom and Yudkowsky, banning systems typically uphold several of the virtues they list. The responsibility of a potential banning is clear, as this rests with the game company in question. Transparency is often substantial as rules of conduct are communicated to the player base, since the game company has no direct interest in banning players. Auditability is enforceable, if by no other means then by collating and examining the ban decisions made by the game company or its model—though whether to release that data or not is typically at the discretion of the company. Incorruptibility is a concern, as players—and in particular players who willfully violate rules of conduct in video games—are often creative and willing to spend time reverse-engineering systems created by game developers. However, there is little risk that the criteria used to decide upon banning would be accidentally changed. Predictability is also usually attainable, again since the game company has an incentive to communicate the criteria that their banning system is enforcing.

Recommender Systems

As more games are released on various platforms each year it becomes progressively harder for players to navigate the full range of available games. Player models may be used as data for curatorial recommender systems that select content on the behalf of players. While this may alleviate the problem of discovering relevant games in a crowded marketplace it also has a number of potential adverse effects for players. Being profiled to a certain kind of games may limit what players are likely to experience. Players who are outliers and do not fit the classes of a given recommender system neatly may receive recommendations that are ill-fitting for them, and they may be less likely to find titles that interest them on their own. On the other side of the relation, game developers who develop niche games may find it harder to reach

wider audiences if their games are modeled as being relevant to a limited subset of profiles.

Aside from the functional aspects of player profiles another issue is that once profiles of individual players are established, the ownership of this profile typically belongs to the individual platform holder. As a piece of information that will influence the player's game experiences going forward, either overtly or covertly, it may be argued from the perspective of Bostrom and Yudkowsky that players should be given options to inspect, impact, or revoke their individual profiles. Relatedly, the profiles driving recommender systems for game content could explicitly or implicitly encode e.g. purchasing power which would factor into recommendations made by the system. This could result in a form of discrimination where more expensive titles are recommended to players with more resources, while marginalized players would only be presented with cheaper titles, reinforcing discrimination in access to varied game experiences.

Recommender systems are usually somewhat transparent, e.g. the Steam platform[1] or Apple's App Store[2] motivate their recommendations with reasons, but the underlying logic driving a recommender system is unavailable to the player. In terms of auditability, recommender systems often are considered trade secrets and are updated so frequently, when new content comes in, that they may be hard to audit at all and also makes them unpredictable. Finally, they are somewhat susceptible to corruption as some developers try to optimize their recommendation rates by exploiting patterns in the systems.

Player Adaptation

Another aspect of the application of models in computer games is the concept of game adaptation. Game adaptation is the tailoring of game content to individual players or clusters of players that individual players are assigned to (Yannakakis and Hallam 2009). When game adaptation is applied, it has the effect of making certain parts of a game's expressive range available to a player, while cordoning off other parts that the player is not shown, due to the adaptation. A commercial example of this is the horror game *Nevermind* (Flying Mollusk 2015) which uses indicators of stress in the player to control the configuration of the game environment. This measurement is in turn used to configure how the game environment is presented and which events occur when in the game. Adaptive experiences like these may tailor the experience to the individual player, making the game more enjoyable, but if the model applied does not correctly capture the particular player, the results may be boring or even unpleasant to the player. More problematically, such features based on physical appearance or physiology may be driven by models optimized for certain subsets of the population. When then imposed on the broader audience it becomes an example of Wallace's type III error. For instance, models have been shown to be unable to properly do facial recognition for persons of certain ethnicities (Eveleth 2016), and virtual reality hardware has been suspected of being optimized for male users at the expense of female users (Boyd 2014; 2000).

Adaptive game play mechanics can have ethical challenges when considered from several of Bostrom and Yudkowsky's virtues. The models are often opaque as machine learning is typically involved for profiling and prediction, which in turn makes internal auditability an issue. The performance of adaptive games can, however, be audited at the player level.

Generative Player Models

Generative player modeling refers to the use of player modeling for representing individual players or clusters of players within games (Holmgård et al. 2015). These methods are still mostly seen within research communities, but instances of these have been seen in commercial game where they enable features such as *ghost play* where an AI agent represents a human player in the game. E.g. games such as *Forza Motorsport 5* (Turn 10 Studios 2013) incorporate partially machine-learned models of individual players called *Drivatars* that represent the human players and function in their place when they are not online and available for play. In these cases, the player model very directly represents its human substrate in the game and becomes responsible for representing that player which may or may not act in accordance with the player's wishes or expectations. This may be problematic if the model does not represent the player accurately. It also raises questions about the ownership of the model that is now a representation of a person much in same way that a photo or a video is, or in this case, a simulation. One specific example may be found in the report of a man using a Drivatar modeled on his father's play style as a memento after the father's death (Riendeau 2014). While the generative player model, according to the news story, was helpful and useful to the player, the Drivatar is wholly owned by the publisher of the game and only available at their discretion. The modeled representation of this man's father is corporate property and he has no immediate options for safe-guarding nor removing this model.

Balancing

Another aspect of computer games strongly impacted by player modeling and profiling is the balancing of games. Balancing typically refers to the adjustment of in-game variables such as the efficacy (damage dealt, speed, health, etc.) and cost (point costs, mana cost, etc.) of in-game entities or abilities. A significant portion of the time spent learning any specific game goes toward understanding the particular interactions between these variables, as they define what strategies are strong and which are weak in most games (Elias, Garfield, and Gutschera 2012). Most multiplayer games that are continuously played via the Internet will usually receive updates to their balance, as the player base evolves and strategies are refined. Choices regarding balancing are typically still made manually by game designers, but are increasingly informed by player models, built from telemetrically collected play data.

[1] http://store.steampowered.com/

[2] http://www.appstore.com

When game designers and analysts change the balance of the game, they are necessarily invalidating aspects of the knowledge of the game's mechanics that players have accrued over time. This serves to keep the game interesting and competitive, but simultaneously nullifies the value of some of the time that players have invested in learning the game and renders unfeasible certain strategies and play styles. In the long term, this favors players who have the time and resources to analyze and experiment with these changes. The models and profiles informing this work indirectly define a threshold for which players can keep up with the game and which players cannot, including some players and excluding others.

Player Matching

In player matching the game suggests an opponent or team the players together. The matching might be aided by requests from the players or entirely based on predictions based on past performance. Prior research has shown how meetings and interactions in online environments such as games can create emotionally meaningful and at times lasting relationships both online and off-line (Taylor 2012). Player matching is typically implemented using player models and profiles that predict the performance of players alone or in groups, such as Elo ratings or match-making ratings. These values may be used to predict the win-probabilities and are used to ensure fairness in matches. The metrics operate only on skill, however, and do not take into account other aspects of players such as consider personality traits or demographic parameters.

Training and Assessment

A challenge for player modeling and profiling that has been emerging over the course of the last decade is the application of games and game-like software for *brain-training* and assessment. Player modeling lies at the core of these applications that claim to either improve or measure players' cognitive abilities. In these cases, the ethical implications of the involved models become comparable to the ones involved in psychological testing. Results obtained from training and testing can influence significant beliefs about own capabilities and may impact workplace decisions such as performance estimations and hiring decisions. Models lacking in reliability or validity may have significant impacts on players' general lives. Here, the need for transparent, auditable, incorruptible, and predictable models is evident, but so far the literature on the efficacy of these approaches is divided on the subject, in particular for brain training (Owen et al. 2010).

Privacy

Modern games have become adept at constantly collecting information about users through telemetry and game analytics. This information can be invaluable to the development and refinement of games over time, showing how different players prefer to interact with a game system or with other players (El-Nasr, Drachen, and Canossa 2013). Additionally, the game industry has seen a number of middleware companies, such as e.g. GameAnalytics[3], make telemetric tools available to game developers of any size, removing the need for a large analytics department before data collection can start.

Canossa (2014) has documented how the existence of these kinds of data collection is seldom communicated clearly to players and if even is is, players usually do not understand or reflect upon this fact. This raises an ethical concern about the privacy of game players and the ownership of the provided data. In-game behavior, including playing style, reactions, in-game choices, naming patterns and many other things correlate significantly with numerous aspects of our lives outside of games. Several studies have shown that it is possible to predict one from the other.

Nick Yee and colleagues investigated how player choices in *World of Warcraft* correlated with the personalities of players. They used data about players' characters from the World of Warcraft Armory and correlated this information with personality tests administered to players; multiple strong correlations were found (Yee et al. 2011). In a similar vein, Canossa et al. investigated how players' life motives correlated with their *Minecraft* log files (Canossa, Martinez, and Togelius 2013). This research used the life motivation questionnaires of Steven Reiss, and found that players' self-reported life motives (independence, family etc) were expressed in a multitude of ways inside constructed Minecraft worlds. Using a very different type of game, Tekofsky et al. have been able to show strong correlations between playing style in the first-person shooter *Battlefield 3* and player characteristics such as personality (Tekofsky et al. 2013), age (Tekofsky et al. 2015) and nationality (Bialas, Tekofsky, and Spronck 2014).

It thus seems that we can make inferences about many different aspects of real-life game players based on their in-game behavior. We do not currently know the limit of what can be learned, or how reliable predictions can be made. But there is no reason to believe that one could not build predictive models of the most sensitive types of information, such as sexual preferences, political views, health status and religious beliefs. One way of seeing this is that when we play a game, we are constantly providing information about ourselves via the controller, keyboard, camera or whatever input device we are using. This is particularly troublesome as we do not perceive ourselves as providing this information in the same way as we would when writing an email or typing into a search box. This is probably partly because we believe ourselves to be "inside the magic circle", and partly because we do not realize how much can be inferred from seemingly innocuous actions. Misuse of this information is a potentially serious problem, given that players do not even know that they can be handing over potential information.

As data mining and modeling techniques improve, it is possible that future work will enable the modeling and profiling of additional preferences, states, and traits from the same existing data sets, yet it is not clear who has the ownership of player data collected during game play and if players

[3] http://www.gameanalytics.com/

can demand their data returned if they no longer wish to part of a company's telemetric database. Reportedly, start-ups are already data mining social media profiles to determine credit scores and assess loan applications (Eichelberger 2013) so it is not inconceivable that data from games could be used in a similar fashion. In situations like this the questions of transparency, audibility, incorruptibility, predictability and minimizing errors become critical.

Methods for Ethical Player Modeling

Player modeling is a useful tool in game development and production and will likely remain a part of computer games in the future. The sections above have outlined some of the general ethical concerns that may arise when applying player modeling and profiling to games while also suggesting a number of specific problems that are present in current game development practices. This raises the question of what methods are available to prevent or address ethical challenges when applying player modeling and profiling.

Several initiatives to promote and support ethical applications of AI have been proposed in recent years, including by important organizations such as IEEE (The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems 2016). However, for many of the examples listed it is reasonable to suggest that game developers and AI modelers from a business perspective may have little interest in following the virtues described in Bostrom and Yudkowsky (2014). This points to the need for external mechanisms regulating the practice, limiting unethical applications. Player modeling that may result in discriminatory actions is likely already illegal, but due to the black box nature of player modeling it can be difficult to detect and police.

For game developers, designers and publishers that choose to care about the questions described in this paper there are several recommendations available and embryos of best practices. We can draw on work ethics for AI as well as in design studies and sociology. A first step would be to increase the diversity of teams researching and implementing player modeling and profiling methods. As models are typically seen to incorporate the values of the people building them, intentionally or unintentionally, diverse teams would be a first preventative measure against biased or discriminating models.

Before starting to develop a model, it is important to think about the potential ways it could be misused or misinterpreted. Essentially, "what could possibly go wrong".

There is a lot to learn from fields such as qualitative sociology and design studies, developed over decades of practice. For example, user-centered design methods such as personas (Pruitt and Grudin 2003) have been used to broaden the understandings of and perspective on the potential and consequences of design artifacts and systems and could most likely yield a deeper understanding of the basis for and impacts of player models and profiles. This helps identifying potentially undesirable consequences of models that implementing teams might not be able to imagine on their own.

To the extent possible, models should be made transparent and explained to players. Most importantly, players deserve to know that modeling is happening and what is being modeled.

Finally, the issue of biased datasets is thorny and not fundamentally solvable. Models and profiles should be investigated and audited continuously, in order to vet them for biased outcomes and qualitatively undesired results. For instance, recent research has shown how machine learned models may be improved in terms of accuracy by seeding them with qualitative evaluations through partially synthetic data sets (Agarwal et al. 2016). In the same vein, research has documented how control data sets may be used to ensure that that outcomes are not based on latent variables that might not be represented in training data sets, but might still be learned by models, such as social class, sex, or ethnicity.

Conclusion

In this paper we raised the question of how we can ensure the ethical application of player modeling and profiling in computer games. Building on prior work in ethics for AI and the ethics of computer games, we presented a number of specific examples of how player modeling and profiling could give cause for ethical concern. Based on these examples, we outlined four general paths forward, focused on diversity in research and implementation, a critical stance toward empirical data, user-centered and participatory model design, and the continuous vetting of models using synthetic data and control data. Identifying these ethical issues and methods useful for addressing them constitutes a first step toward the ethical practice of player modeling. Ensuring incentives for adhering to these principles for researchers and industry is a separate problem that most likely requires a combination of community standards, codes of conduct, and outside regulation and legislation.

References

Agarwal, A.; Bird, S.; Cozowicz, M.; Hoang, L.; Langford, J.; Lee, S.; Li, J.; Melamed, D.; Oshri, G.; Ribas, O.; et al. 2016. A multiworld testing decision service. *arXiv preprint arXiv:1606.03966*.

Alha, K.; Koskinen, E.; Paavilainen, J.; Hamari, J.; and Kinnunen, J. 2014. Free-to-play games: Professionals' perspectives. *Proceedings of Nordic Digra* 2014.

Bialas, M.; Tekofsky, S.; and Spronck, P. 2014. Cultural influences on play style. In *Computational Intelligence and Games*, 1–7. IEEE.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Bostrom, N., and Yudkowsky, E. 2014. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence* 316–334.

Boyd, D. 2000. Depth cues in virtual reality and real world: Understanding individual differences in depth perception by studying shape-from-shading and motion parallax. *BA Thesis, Department of Computer Science, Brown University, Providence RI*.

Boyd, D. 2014. Is the oculus rift sexist? *Quartz*.

Canossa, A.; Martinez, J. B.; and Togelius, J. 2013. Give me a reason to dig minecraft and psychology of motivation. In *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*, 1–8. IEEE.

Canossa, A. 2014. Reporting from the snooping trenches: Changes in attitudes and perceptions towards behavior tracking in digital games. *Surveillance & Society* 12(3):433.

Castronova, E. 2001. Virtual worlds: A first-hand account of market and society on the cyberian frontier. *CESifo working paper series*.

Crawford, K.; Whittaker, M.; Elish, M. C.; Barocas, S.; Plasek, A.; and Ferryman, K. 2016. The ai now report: The social and economic implications of artificial intelligence technologies in the near-term.

Drachen, A.; Thurau, C.; Togelius, J.; Yannakakis, G. N.; and Bauckhage, C. 2013. Game data mining. In *Game Analytics*. Springer. 205–253.

Eichelberger, E. 2013. Your deadbeat facebook friends could cost you a loan. *Mother Jones*.

El-Nasr, M. S.; Drachen, A.; and Canossa, A. 2013. *Game Analytics: Maximizing the Value of Player Data*. Springer Science & Business Media.

Elias, G. S.; Garfield, R.; and Gutschera, K. R. 2012. *Characteristics of Games*. MIT Press.

Eveleth, R. 2016. The inherent bias of facial recognition. *Motherboard*.

Flying Mollusk. 2015. Nevermind.

Gee, J. P. 2005. Learning by design: Good video games as learning machines. *E-Learning and Digital Media* 2(1):5–16.

Holmgård, C.; Liapis, A.; Togelius, J.; and Yannakakis, G. N. 2015. Evolving models of player decision making: Personas versus clones. *Entertainment Computing*.

Holmgård, C.; Togelius, J.; and Henriksen, L. 2016. Computational intelligence and cognitive performance assessment games. In *Computational Intelligence and Games (CIG)*.

Huizinga, J. 1955. *Homo Ludens: A Study of the Play-element in Culture*. Beacon Press.

Kato, P. M. 2010. Video games in health care: Closing the gap. *Review of General Psychology* 14(2):113.

Larsson, T., and Dodig-Crnkovic, G. 2004. Game ethics-homo ludens as a computer game designer and consumer. *Game Studies* 4(1).

Levin, S. 2016. A beauty contest was judged by ai and the robots didn't like dark skin. *The Guardian*.

Lum, K., and Isaac, W. 2016. To predict and serve? *Significance* 13(5):14–19.

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group (NY).

Owen, A. M.; Hampshire, A.; Grahn, J. A.; Stenton, R.; Dajani, S.; Burns, A. S.; Howard, R. J.; and Ballard, C. G. 2010. Putting brain training to the test. *Nature* 465(7299):775–778.

Pearce, C.; Boellstorff, T.; and Nardi, B. A. 2011. *Communities of play: Emergent cultures in multiplayer games and virtual worlds*. MIT Press.

Pruitt, J., and Grudin, J. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*, 1–15. ACM.

Riendeau, D. 2014. Son races deceased father's ghost in Xbox racer, makes internet cry. *Polygon*.

Salen, K., and Zimmerman, E. 2004. *Rules of play: Game design fundamentals*. MIT press.

Shute, V., and Ventura, M. 2013. *Stealth assessment: Measuring and supporting learning in video games*. MIT Press.

Sicart, M. 2011. *The ethics of computer games*. MIT Press.

Smith, A. M.; Lewis, C.; Hullett, K.; Smith, G.; and Sullivan, A. 2011. An Inclusive Taxonomy of Player Modeling. *University of California, Santa Cruz, Tech. Rep. UCSC-SOE-11-13*.

Sweeney, L. 2013. Discrimination in online ad delivery. *Queue* 11(3):10.

Taylor, T. L. 2009. *Play between worlds: Exploring online game culture*. MIT Press.

Taylor, T. 2012. *Raising the Stakes: E-sports and the Professionalization of Computer Gaming*. MIT Press.

Tekofsky, S.; Spronck, P.; Plaat, A.; Van den Herik, J.; and Broersen, J. 2013. Psyops: Personality assessment through gaming behavior. In *Benelux Conference on Artificial Intelligence*.

Tekofsky, S.; Spronck, P.; Goudbeek, M.; Plaat, A.; and van den Herik, J. 2015. Past our prime: A study of age and play style development in battlefield 3. *IEEE Transactions on Computational Intelligence and AI in Games* 7(3):292–303.

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems*. IEEE.

Turn 10 Studios. 2013. Forza Motorsport 5.

Wallace, W. A. 1994. Ethics in modeling.

Woodford, D. 2008. Abandoning the magic circle. In *Breaking the Magic Circle, Game Research Lab Seminar, University of Tampere*.

Yannakakis, G. N., and Hallam, J. 2009. Real-time game adaptation for optimizing player satisfaction. *IEEE Transactions on Computational Intelligence and AI in Games* 1(2):121–133.

Yannakakis, G. N.; Spronck, P.; Loiacono, D.; and André, E. 2013. Player Modeling. In *Artificial and Computational Intelligence in Games*. Saarbrücken/Wadern: Dagstuhl Publishing. 45–55.

Yee, N.; Ducheneaut, N.; Nelson, L.; and Likarish, P. 2011. Introverted elves & conscientious gnomes: The expression of personality in world of warcraft. In *CHI*, 753–762. ACM.

Zagal, J. P.; Björk, S.; and Lewis, C. 2013. Dark patterns in the design of games. In *Foundations of Digital Games*.