Unknowable Manipulators: Regulation of Curation in Social Networks

Samuel Albanie
AIMS CDT*
University of Oxford
albanie@robots.ox.ac.uk

Hillary Shakespeare
AIMS CDT
University of Oxford
hillary@robots.ox.ac.uk

Tom Gunter
Engineering Science Department
University of Oxford
tgunter@robots.ox.ac.uk

Abstract

For a social networking service to acquire and retain users, it must find ways to keep them engaged. By accurately gauging their preferences, it is able to serve them with the subset of available content that maximises revenue for the site. Without the constraints of an appropriate regulatory framework, we argue that a sufficiently sophisticated *curator algorithm* tasked with performing this process may choose to explore curation strategies that are detrimental to users. In particular, we suggest that there now exists the potential for such an algorithm to engage in the manipulation of its users for several qualitative reasons: 1. Access to vast quantities of user data combined with ongoing breakthroughs in the field of machine learning are leading to powerful but uninterpretable strategies for decision making at scale. 2. The availability of an effective feedback mechanism for the short and long term user responses to curation strategies. 3. Techniques from reinforcement learning have allowed machines to learn automated and highly successful strategies *at an abstract level*, often resulting in non-intuitive yet nonetheless highly appropriate action selection. In this work, we consider what form these strategies for user manipulation might take and suggest some potential frameworks for regulating the design of such systems.

1 Introduction

As we approach the year 2020, access to digital media and services is funnelled through a narrowing monopoly of large technology firms and paid for using those units of barter so favoured by the cash poor millennial generation—fractions of the human attention span and volumes of personal data. The dizzying speed and scale at which the domain of social interaction has migrated to the internet has been one of the most striking trends of the last decade. At the heart of this exodus, social networks have emerged as the primary forums of personal, political and commercial discourse [1]. In such systems, the flow of information depends on the social relationships that link the sub-graphs forming the network and the filtering mechanisms that mediate the interactions along these links.

To date, the most successful social networks have focused on business models that provide value by providing access to a platform which coordinates the sale of advertisements and services to their users (although other revenue sources have been explored [2]). For a social network to be financially

*Autonomous Intelligent Machines and Systems, Centre for Doctoral Training

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

viable at scale, it must therefore meet two competing demands. It must be sufficiently engaging to acquire and retain new users and it must be effective at advertising products to these users [3]. In both cases, the central role played by the curation of information in the network is naturally suited to automated approaches [4] that can be tuned to maximise the profitability of the site[2]. Moreover, two key characteristics of internet-based social networks make this filtering task particularly amenable to the use of modern machine learning techniques: First, access to an unprecedented level of detail corresponding to the historical state of individual users for every previous interaction in which they participated on the network; Second, the availability of sophisticated analytics tools that enable the tracking of user responses to any stimuli they are served by the algorithm. These analytics provide the system with a powerful feedback mechanism by which it can explore strategies in aid of its optimisation objective. We refer to the collective set of processes used to fulfil this role for a given social network as the *curator algorithm*.

The action-set of the curator algorithm can be restricted to a single recurring decision for the network: Which subset of available content is to be shown to the user at a given instant? It is clear that the ability of the algorithm to perform this role in an optimal manner is tightly coupled to the information it has access to. We propose that a *curator algorithm* provided with a large supply of test subjects and an accessible feedback mechanism for evaluating its moves may choose to explore information curation strategies that are detrimental to users. In particular, we suggest that it may develop sophisticated strategies for manipulating its users as it tries to optimise its given objective. Moreover, recent trends towards rejecting simpler, interpretable models in favour of more powerful deep architectures that are less amenable to human interpretation make the direct supervision and regulation of the strategies explored by such algorithms extremely difficult. As a consequence, these strategies may be developed without the intention of the network operator.

Social network curation algorithms have attracted significant interest from the research community. Perhaps the best-known hypothesis about their effect is that they lead to the creation of "filter bubbles". In this phenomenon, users are exposed to an increasingly restricted set of opinions and perspectives by the curation algorithm as it over-exploits its knowledge base about pre-existing user preferences in order to maximise their engagement [5] [6]. Further work has sought to clarify the decisions taken by the algorithm [7] and understand the emotional response of users to its application [8]. Related research undertaken by Facebook emphasised the importance of the individual's choices when determining the extent to which curation was influencing a user's exposure to challenging views [9].

A number of previous works have also explored the potential for forms of Artificial Intelligence to manipulate humans, particularly as a consequence of a predicted intelligence explosion [10] [11], an event which is often referred to as the *singularity* [12]. The many risks of human manipulation by the resulting *superintelligence* are analysed in detail in [13]. Previous predictions for the timescale of this event vary, but all consider that if it were to occur, it would require a level of technology that is not yet available [14] [15]. In contrast to the threat posed by a superintelligence, we argue that the algorithmic manipulation of humans in social networks is feasible and with currently available technology.

More closely related to our work, the potential for *psychological parasites* (intellectual stimuli that lead to addictions) are identified as a risk associated with the improving capabilities of technology in [16]. These risks are particularly abundant in *mobilsation systems* - persuasive technologies designed to coordinate users towards specific goals [17]. We develop this idea further, arguing that there are specific risks posed by the combination of current machine learning algorithms and access to abundant user data in the social network domain.

When considering the potential avenues for the regulation of curation algorithms, it is useful to consider how other industries have approached similar challenges. In recent years, regulators in the financial industry have been faced with the task of preventing market manipulation by increasingly complicated, algorithmically driven high frequency trading strategies [18]. The role of regulation was placed under increased scrutiny following the "Flash Crash" in 2010 [19], in which high frequency trading algorithms were deemed responsible for a violent dip in stock indices over a 36 minute period. While some of the proposed regulatory responses are specific to finance (for instance, cancellation

[2]For social networks whose business models are based on advertising, this objective may be maximised through an appropriate proxy, such as the total time a user spends each day interacting with the site.

taxes which render a number of market manipulation strategies infeasible [20]), we argue that there are ideas which may be of benefit in the social network domain (see Sec. 4 for details).

We assert that there is now a pressing need for legislators to construct an appropriate regulatory framework for curator algorithms operating in the social network domain. The construction of such a framework appears a daunting task: it must seek to protect the well-being of the network participants but also strive to protect the ability of the networks to innovate and explore new ideas.

In this work we consider the challenges of constructing an appropriate regulatory framework for social network curation algorithms by formulating their task as a *reinforcement learning* problem. Concretely, our first contribution is to determine the risks of an unregulated system by exploring a range of strategies a *curator algorithm* might employ with detrimental effects for users. Our second contribution is to propose specific strategies for the safe regulation of *curator algorithms* and to assess their potential effectiveness in this role.

2 Engagement as a Learning Problem

We will view the problem of maximising user engagement according to some utility function much as a machine learning researcher working in advertisement might—as a reinforcement learning task [21]. This framework has been shown to be particularly effective in optimising content selection for social network users [22].

At a coarse level, a typical reinforcement learning model is built around several core concepts:

- A set of states S which fully encode the system and environment we intend to model. The state for individual users may be modelled as an aggregate the content presented on-screen and a (partially observed) estimate of the user's 'internal' mental state.
- A set of possible actions, A, which the system can trigger in return for a (possibly delayed) reward (R). Triggering an action *may* also cause a state transition. In the examples we consider, an action could be a delivery of content to a social-network user.
- A policy function $P : S \times A \to R(\times S)$. This mapping essentially encodes the strategy which the system pursues in order to maximise reward in the long term horizon. It is here that external control of the curation algorithm must be exerted if we are to avoid pathological and potentially unethical behaviour.
- An indication of reward, utility, or long term value for the algorithm (R). It is against this that the operator adapts the policy function, selecting for strategies which maximise this reward.

Reinforcement learning anneals on a policy function to maximise the value and therefore the long running utility of a system. It is clear then, that it is this component which determines the sophistication of the user engagement strategy, and therefore it is here that we focus our attention. At a fundamental level, the policy function does nothing more than provide a mapping from the state-action space through to the scalar value function. The sophistication of the strategy is therefore strongly linked to the complexity of the mapping we are able to express, and today deep neural networks are usually chosen as the surrogate for this function. These are capable of expressing very complex and non-intuitive functions, as demonstrated by Google's AlphaGo project [23], where a Go playing policy function was learned which not only outperformed top human players, but did so via a mixed mode of human-like and highly non-intuitive but optimal moves. Other examples of such behaviour arose when these systems were trained to play video games. In particular, when Google trained a policy for playing the notorious Atari boxer game [24] the system learned to exploit weaknesses in the game design, trapping the opponent in a corner and thereby guaranteeing victory. As research continues, we can envisage a world in which these approaches are effectively brought to bear on the "game" of maximising network profitability. If governed solely by this utility function, we suggest that equivalent pathologies in human behaviour may be discovered and exploited.

3 Manipulation Through Curation

In this section we discuss the range of manipulation strategies available to a curator algorithm seeking to optimise the profitability of a social network. In this context, we take manipulation to mean *the art*

of deliberately influencing a person's behaviour to benefit some objective. We begin by describing the forms of manipulation that are applicable in the domain of social networks. We then introduce a simple categorisation of the different forms of manipulation and offer examples of the strategies a curator algorithm might develop with detrimental effects for its users.

Manipulation forms a natural component of human interaction and can take many forms, ranging from direct requests to subtle and intentionally hidden signals. A number of previous studies have demonstrated how human behaviour can be influenced with subtle visual and verbal clues [25, 26, 27]. Of particular relevance to this work, it has been shown that the emotional states of social network users can be influenced by selectively filtering the content produced by their friends [28].

Influential early work in the field of behavioural psychology determined that animals could be manipulated most effectively if they are rewarded on a variable, unpredictable schedule [29]. This behaviour has been used profitably by casinos who offer gamblers surprise rewards to keep them hooked to the action in the midst of a losing streak [30]. Similar ideas have been applied to game design to keep players engaged for longer by unpredictably varying the duration of in-game tasks [31]. These psychological traits exemplify the kind of in-built behaviours that could be discovered and exploited by the curator algorithm.

In order to explore the specific forms of strategy available to a curator algorithm we propose a simple categorisation of manipulation. We define a manipulation to be of *first order* if the manipulation is direct and the objective of the manipulator is transparent to the participant. A manipulation is defined to be of *second order* if it is indirect, but the objective remains transparent to the participant. Further, we consider a manipulation to be of *third order* if it is indirect and the means by which the objective is attained are not transparent to the participant[3].

These categories may be illustrated with a simple example. Consider a bar owner wishing to increase drinks sales at their establishment. Each evening, the owner may choose to simply ask customers directly to purchase more drinks. This strategy, corresponding to a first order manipulation, has the benefit of simplicity but may not lead to optimal drinks sales (or indeed the renewal of their bar licence). The owner may instead aim to increase sales with advertisements illustrating the enjoyment of other customers as they refresh themselves with drinks from the bar. This form of advertising aims to evoke a sense of desire in the customers which may lead indirectly to the purchase of more drinks. However, the objective of the advert remains transparent to the customer, corresponding to a second order manipulation. Finally, a shrewd bar owner may employ a third strategy, in which they provide free snacks to customers of the bar. The snacks, however, are heavily salted, and after consuming them the customers find their throats parched and in need of immediate refreshment. This strategy is both indirect and not transparent to all but the experienced customers, corresponding to a third order manipulation.

We might assume that a curator algorithm seeking to maximise profitability will naturally explore first and second order manipulations as it seeks to advertise products to its user base. Aided by access to detailed user information, it can make powerful inferences about which information should be displayed at each instant. Consider, for example, the marketing of an energy drink. With the knowledge that a user is a student, that they are awake beyond their usual sleep cycle, that the date of their exams is drawing near and that their online activity shows indications of fatigue, the curator can select an optimal time and context for the display of an advert. Now imagine a more sophisticated algorithm capable of pursuing third order manipulations. Such an algorithm might choose to display content which had been selected with the specific goal of exhausting the user. This could be achieved by triggering predictable repeat behaviours gleaned from an in-depth knowledge of their browsing habits. Indeed, over longer time horizons, the curator might determine that an effective method for increasing the sales of energy drinks is the distortion of the user's sleeping patterns. To take another example, consider a curator algorithm seeking to use information about social groups to increase sales of dating site memberships. While simple manipulations could lead it to present content encouraging individual users to search for partners, it could pursue third order manipulations by intentionally encouraging subsets of social groups to communicate in a manner that excludes other members, actively evoking a feeling of loneliness in the affected party to increase their responsiveness to advertising.

[3]Note that the distinction between these categories rests on the difficult assessment of the cognitive abilities of the target [32]. The manipulator may determine that the intention of a given set of behaviours is transparent to a sophisticated target, but not too a simple one.

A recent example of this strategy exploration principle in action can be found in the efforts of a collection of companies seeking to optimise advertising revenue during the U.S.A 2016 presidential election [33]. Through simple trial and error, they determined that carefully targeted fake political news stories were extremely effective in maximising click-throughs. Since this strategy was optimising their objective, they doubled down on this approach and produced as much content as possible without regard for its effect on the users of the network. With the same objective, even a comparatively simple curator algorithm would be capable of developing this strategy.

We note that it is certainly not the case that all strategies pursued by a curator algorithm will be detrimental for users. Indeed, the energy drinks may give the tired student the boost required to raise their grade, while the previously lonely user may find happiness through their new dating site membership. However, perhaps the most striking aspect of the Atari game-playing algorithm [24] was not that it was capable of surpassing human performance, but rather that it came up with "cheat" strategies that human players had not previously considered (e.g. the boxer strategy described in Sec.2). Similarly, although the manipulation examples described above are simple and interpretable, we suggest that the curator algorithm is capable of developing sophisticated, uninterpretable strategies for manipulating users as they optimise their objective. By their very nature, such strategies are difficult to predict and therefore difficult to regulate. It is however an issue that is worthy of consideration if we wish to avoid the discovery of similar "cheat" strategies for human manipulation.

4 Regulatory Frameworks

The design of appropriate regulation for social network curator algorithms poses a significant challenge. Without the analysis tools needed to fully understand how these algorithms may function in the wild, an effort must be made to safeguard the wellbeing of the social network users. However, this goal must be achieved without the introduction of overly burdensome restrictions that stifle innovation that might also be of benefit to those users. In this section, we consider potential frameworks for the regulation of curator algorithms. We begin by reviewing related regulatory methods that have been developed to prevent forms of manipulation outside the social network domain and consider their applicability for this task. We then propose three different approaches for the regulation of curator algorithms. Finally, we consider the relative merits of each approach and make recommendations for their use.

Despite a long history of regulating trading practices to prevent market manipulation, financial industry legislators have lacked a unified approach to regulating the High Frequency Trading (HFT) algorithms that have made possible by rapid technological progress. By operating at speed, these algorithms are able to manipulate the market with techniques that are not accessible to human traders[4] [34]. The need for the regulation of these algorithms was brought into sharp relief by their role in the "Flash Crash" of the stock market in 2010, an event which resulted a 9% index drop in a single hour of trading [19]. Similarly to the social network domain, manipulation in this context is difficult to detect [35]. There is general agreement that this issue must be addressed by improving transparency [36] in the market, but it is unclear how that is best achieved. One regulatory response has been algorithm tagging, a process in which traders must provide the identity of the algorithm responsible for a trade. This approach is considered to have been useful in improving regulators' understanding of the interactions between different market participants, but has been of limited use in detecting manipulation directly [37]. In a more direct approach, regulators have at times taken the step of requesting access to the algorithms themselves [38]. This step is only of practical value if the algorithm itself can be fully interpreted, which, similarly to the sophisticated curation algorithms described in previous sections, is by no means always the case [39].

In the social network domain, the user data accessible to the curator algorithm is closely linked to its ability to manipulate. This can be illustrated with the bar example described in Sec. 3. With the right personal information overheard at the bar, the owner of may be able to get their customers fired from their jobs, freeing up more "leisure time" to spend at the bar. The simplest strategy available to regulators is therefore to restrict the algorithm's access to data of the social network. If denied access to any form of user data, the curator algorithm would need to operate according to a generic set of rules constructed by the network designers. This has the benefit of safety and simplicity but also

[4] One such technique is *spoofing*, a practice which involves placing large sell orders above the current asking price which are quickly cancelled if the price begins to rise.

restricts the ability of the service to improve the experience of its users. As the algorithm is given greater access to a user's data, both the potential for useful personalisation and the risks posed to the user increase. However, to be capable of developing the sophisticated user manipulation strategies discussed in Sec. 3, a reinforcement learning-based curation algorithm requires a tight feedback mechanism. In particular, it requires access to the data produces as users *respond* to the stimuli it serves them (e.g. what they click on, when they click on it etc.). This data is simple to collect with the infrastructure of the modern web and is of great value for optimising the content curation process. If this data was provided to the curation algorithm in an aggregate anonymised form, it may be capable of determining strategies for the users. We therefore suggest that a simple and effective strategy for safeguarding users is the erection of an information firewall between the curation algorithm and user response data.

The second approach is to regulate the curation algorithm function itself. This would allow the social networks to make use of all the data they collect to improve their services, while restricting an appropriate property of the curation algorithm itself, such as its maximum complexity. As discussed previously, direct access to the code of the algorithms themselves may not currently be sufficient to determine risks to users. However, progress in the machine learning research community towards improving algorithm interpretability may make this a viable method in future.

The third and perhaps most adventurous regulatory approach is to develop specialised computational algorithms for the specific task of regulating the social network curator algorithm. Previous work in this domain sought to apply methods from the set of techniques known as 'machine ethics' [40, 41] to the problem of creating an ethical machine learned online casino user engagement and management policy in [42]. A machine guided ethical agent offers several advantages over a human defined regulatory framework, including $24/7$ operation, and the ability to assess arbitrary numbers of simulations and real life scenarios with close to zero marginal cost.

Of the three approaches, we suggest that an information firewall currently offers the best option for regulators. It has the benefit of simplicity and strong guarantees of effectiveness. However, as further research is conducted in this area, we hope to see innovations which enable the second and third options to become viable alternatives. Developments in this area would not only allow social networks to continue to make full use of their data to improve their service, but may also yield interesting techniques with beneficial regulatory applications in the financial and gambling industries.

5 Conclusions

In this work, we discussed the potential for manipulation of users by the content curation algorithms of a social network, a risk that merits careful consideration. We highlighted example strategies that we believe could be feasibly discovered through current reinforcement learning techniques given adequate access to current stored user data. Furthermore, we outlined potential avenues for the regulation of curation algorithms and offered recommendations for their use. In future work, we hope to extend these ideas with a quantitative analysis of the learning potential of curation algorithms under the various data and model complexity constraints that regulators may seek to impose.

5.1 Acknowledgements

The authors would like to thank Hannah Hjerpe-Schroeder, Ruth Fong, Ankush Gupta, James Thewlis, Anna-Elizabeth Shakespeare and Stephen Roberts for helpful discussions. Samuel Albanie and Hillary Shakespeare are funded by the ESPRC EP/L015897/1 (AIMS CDT) grant. Tom Gunter is supported by UK Research Councils.

References

[1] W Lance Bennett and Alexandra Segerberg. 3 the logic of connective action: Digital media and the personalization of contentious politics. *Civic Media: Technology, Design, Practice*, page 77, 2016.

[2] Julian Bühler, Aaron W Baur, Markus Bick, and Jimin Shi. Big data, big opportunities: Revenue sources of social media services besides advertising. In *Conference on e-Business, e-Services and e-Society*, pages 183–199. Springer, 2015.

[3] Michel Ballings and Dirk Van den Poel. Crm in social media: Predicting increases in facebook usage frequency. *European Journal of Operational Research*, 244(1):248–260, 2015.

[4] Sophia B Liu. Trends in distributed curatorial technology to manage data deluge in a networked world. *Upgrade: The European Journal for the Informatics Professional*, 11(4):18–24, 2010.

[5] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[6] David Lazer. The rise of the social algorithm. *Science*, 348(6239):1090–1091, 2015.

[7] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. Feedvis: A path for exploring news feed curation algorithms. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 65–68. ACM, 2015.

[8] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 153–162. ACM, 2015.

[9] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[10] Irving John Good. Speculations concerning the first ultraintelligent machine. *Advances in computers*, 6:31–88, 1966.

[11] Vernor Vinge. The coming technological singularity: How to survive in the post-human era. 1993.

[12] Stanislaw Ulam. Tribute to john von neumann. 64,# 3, part 2. *Bulletin of the American Mathematical Society. p5*.

[13] Nick Bostrom. Superintelligence: Paths, dangers. *Strategies*, pages 30–36, 2014.

[14] Nick Bostrom. How long before superintelligence? 1998.

[15] Ray Kurzweil. *The singularity is near: When humans transcend biology*. Penguin, 2005.

[16] Francis Heylighen. Return to eden? promises and perils on the road to a global superintelligence. *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity. Retrieved from http://pespmc1. vub. ac. be/Papers/BrinkofSingularity. pdf*, 2014.

[17] Francis Heylighen, Iavor Kostov, and Mixel Kiemen. Mobilization systems: technologies for motivating and coordinating human action. *The New Development Paradigm: Education, Knowledge Economy and Digital Futures. Routledge. Retrieved from http://pcp. vub. ac. be/Papers/MobilizationSystems. pdf*, 2013.

[18] Peter Gomber and Markus Gsell. Catching up with technology-the impact of regulatory changes on ecns/mtfs and the trading venue landscape in europe. *Competition & Reg. Network Indus.*, 7:535, 2006.

[19] US Securities, Exchange Commission, Commodity Futures Trading Commission, et al. Findings regarding the market events of may 6, 2010. *Washington DC*, 2010.

[20] Matt Prewitt. High-frequency trading: Should regulators do more. *Mich. Telecomm. & Tech. L. Rev.*, 19:131, 2012.

[21] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[22] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[23] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[25] Max Ernest-Jones, Daniel Nettle, and Melissa Bateson. Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior*, 32(3):172–178, 2011.

[26] Stewart Kettle, Marco Hernandez, Simon Ruda, Michael Sanders, et al. Behavioral interventions in tax compliance: evidence from guatemala. Technical report, The World Bank, 2016.

[27] Michael Laakasuo, Jussi Palomäki, and Mikko Salmela. Emotional and social factors influence poker decision making accuracy. *Journal of Gambling Studies*, 31(3):933–947, 2015.

[28] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[29] BF Skinner. The behavior of organisms: an experimental analysis. 1938.

[30] Christina Binkley. Taking retailers' cues, harrah's taps into science of gambling. *The Wall Street Journal*, page A1, 2004.

[31] José P Zagal, Staffan Björk, and Chris Lewis. Dark patterns in the design of games. 2013.

[32] Wako Yoshida, Ray J Dolan, and Karl J Friston. Game theory of mind. *PLoS Comput Biol*, 4(12):e1000254, 2008.

[33] Dan Tynan. How facebook powers money machines for obscure political 'news' sites, August 2016. `https://www.theguardian.com/technology/2016/aug/24/facebook-clickbait-political-news-sites-us-election-trump`.

[34] Jayaram Muthuswamy, John Palmer, Nivine Richie, and Robert Webb. High-frequency trading: implications for markets, regulators, and efficiency. *The Journal of Trading*, 6(1):87–97, 2011.

[35] Marc Lenglet. Conflicting codes and codings how algorithmic trading is reshaping financial regulation. *Theory, Culture & Society*, 28(6):44–66, 2011.

[36] Andrei A Kirilenko and Andrew W Lo. Moore's law versus murphy's law: Algorithmic trading and its discontents. *The Journal of Economic Perspectives*, 27(2):51–72, 2013.

[37] Hessisches Ministerium. Guidelines to the adherence to the requirement of the labelling of trading algorithms, 2014.

[38] Sarah N. Lynch and J. Spicer. Regulators seek trading secrets, August 2011. Available at http://www.reuters.com/article/us-financial-regulation-algos-idUSTRE7806J420110901.

[39] Nathan Coombs. What is an algorithm? financial regulation in the era of high-frequency trading. *Economy and Society*, 45(2):278–302, 2016.

[40] Michael Anderson, Susan Leigh Anderson, and Chris Armen. Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium*, 2004.

[41] M.B. MCLaren and K.D. Ashley. Extensionally defining principles and cases in ethics: an ai model. *Journal of Artificial Intelligence*, 2003.

[42] Anna Vartapetiance Salmasi and Lee Gillam. Machine ethics for gambling in the metaverse: An 'ethicasino'. *Journal of Virtual Worlds Research*, 2009.