The problem of behaviour and preference manipulation in AI systems

Hal Ashton, Matija Franklin

University College London
ucabha5@ucl.ac.uk

Abstract

Statistical AI or Machine learning can be applied to user data in order to understand user preferences in an effort to improve various services. This involves making assumptions about either stated or revealed preferences. Human preferences are susceptible to manipulation and change over time. When iterative AI/ML is applied, it becomes difficult to ascertain whether the system has learned something about its users, whether its users have changed/learned something or whether it has taught its users to behave in a certain way in order to maximise its objective function. This article discusses the relationship between behaviour and preferences in AI/ML, existing mechanisms that manipulate human preferences and behaviour and relates them to the topic of value alignment.

Introduction

Increased data collection possibilities in the modern age mean that Statistical Artificial Intelligence (AI) or Machine Learning (ML) are often used to learn the preferences of users in order to better (sometimes for the user, sometimes to the system owner) deliver some service to them. Preferences can be learned directly by asking subjects directly (Stated Preferences) or they can be inferred in a process known as Revealed Preference Theory (RPT) (Varian 2006). Both approaches come with an extensive set of limitations which have been demonstrated over time by experimental economists and psychologists. One set of limitations broadly falls into the category of 'irrational' behaviour or beliefs. For example Gui, Shanahan, and Tsay-Vogel (2021) discuss the phenomenon of users acting inconsistently as they balance conflicting short and long term preferences. Preferences might not be static between contexts; the social norms of people 'in-group' (Cialdini and Trost 1998), might run contrary to their person's private preference, revealed through their digital behaviour. The presence of multiple preferences active in different circumstances poses the question which preference 'revealed' from behaviour ought to be selected by decision makers as the 'true' preference or 'normative' preferences (Beshears et al. 2008). Decision makers might also make mistakes (Nishimura 2018), be susceptible to various environmental effects like framing (Tversky and Kahneman 1985), and they may exhibit satisficing where users do not even view the best option because of search costs (Caplin, Dean, and Martin 2011).

We will concentrate on a problem with preference elicitation and representation which we argue when combined with the iterative nature of AI/ML risks can cause profound problems. The issue stems from user preferences being quite fluid and changeable in practice (Bleidorn, Hopwood, and Lucas 2018; Mathur, Moschis, and Lee 2003) and worse, they can be influenced in any number of ways. The existence of a large and successful behavioural change industry, with practitioners in government and advertising, is evidence of this. This is relevant to preferences because, amongst others, Ariely and Norton (2008) have shown that behaviour is not only caused by preference but also the inverse is true: Behaviour causes preferences to form.

This article will explore the implications of non-static preferences and plastic behaviour/preferences when AI/ML systems are tasked with learning user preferences over time. It will point to a small but growing body of research that shows that the plasticity of human preferences under algorithmic influence is a profound problem without obvious solutions.

Behaviour change accepted; preference change unacknowledged

There is a large body of research showing that the behaviour of users can be reliably changed with a variety of techniques. The commercial side of this behaviour change complex comprises the advertising industry (Sutherland 2019) and the academic side falls under the umbrella of behavioural science (Ruggeri 2018), typically distributed across but not limited to Business schools, Psychology and Economics departments. The practice was brought to popular attention by Thaler and Sunstein (2008), with the virtuous behaviour change practice called 'nudging'. Specifically this is the development of choice architectures, the background for people's behaviours, aimed at influencing people's behaviour, without limiting or forcing options, or significantly changing their economic incentives. A major consumer of nudging expertise has been governments; to date nudging has been used as a policy tool in over 80 countries and by supranational institutions (OECD 2017).

All environments influence behaviour to some extent, even when people are not aware of it (Sunstein 2016). To give a concrete example, content recommender engines, even if not labelled as such, nudge their users because they deliberately alter the choices that a user can make when delivering personalised search results on the first page of results in web browsers, or projected onto maps in cars and phones, or when suggesting further things to watch on the TV.

The observation that behaviour can be changed by system designers (Schneider, Weinmann, and vom Brocke 2018; Kozyreva, Lewandowsky, and Hertwig 2020) through changes in choice architectures or other techniques immediately calls into question the practicality of user preference elicitation and in particular RPT. This is because there is a considerable body of evidence showing that behaviour history forms preferences (Ariely and Norton 2008; Albarracín and Jr 2000; Albarracín and McNatt 2005; Hill, Kusev, and van Schaik 2019; Wyer, Xu, and Shen 2012). A response might be to say that behaviour which has been altered does not reflect the ’real’ or ’normative’ preferences of a user and better efforts should be made to learn un-manipulated preferences. Firstly this is not trivial for any preference learner because it means they then have to distinguish between representative and non-representative behaviours in their data. Secondly it is naïve because it does not allow users to autonomously change their preferences (by developing a taste for Nollywood cinema or Mongolian throat singing say).

The behaviour change complex overcomes the difficulty in eliciting preferences by not really modelling them; behaviour is the key metric of success (Atkins et al. 2017). Whether someone who has had their behaviour changed prefers their new behaviour to their old one is not usually a focus. Proponents of the use of behaviour change defend the practice ethically by arguing that they only influence behaviour, but do not limit or force options. This is described as Libertarian Paternalism, a form of soft means paternalism, with the central idea that institutions can positively affect people’s behaviour, while still respecting their freedom of choice (Thaler and Sunstein 2003). It is described as ‘soft’ because it avoids material incentives and coercion, thus maintaining freedom of choice; and as ‘means-orientated’ because it does not attempt to change people’s goals (or ends), but rather gives people a sense of best practice, given their own ends. Proponents of libertarian paternalism favour the intentional design of choice architecture as a policy tool (Sunstein 2014). They argue that since choice architecture is omnipresent, unavoidable and influences people’s behaviour, even when they are not aware of it, so it might as well be harnessed to do good. Nevertheless the Libertarian Paternalist argument seldom considers the observation that behaviour change has a causal relationship with preference change.

Private sector companies are wary about stating an intent to change user behaviour because of the likely public opprobrium which may occur. This attested by the recent popularity of popular media examining the manipulative behaviour of big-tech attest (Orlowski, Coombe, and Curtis 2020). As a result of the public’s sensitivity surrounding behaviour change, the objective of behaviour change is couched in terms of preference learning - the desire to learn about customers to better engage with them and improve their user experience. On the occasions that companies have been shown to use AI to maximise profitable behaviour over maximising an objective function based on user preferences, the public reception has not been warm (Lewis and McCormick 2018). Training a video recommender to maximise play-through because more complete videos watched equals to more adverts consumed fulfils a logical business objective but in the language of behavioural change, has spillovers (Dolan and Galizzi 2015). As Alfano et al. (2020) show, such a system can involve recommending extremist content to maintain users’ attention. Ignoring preference change in this case ignores the social externality that AI/ML powered behaviour change causes. The impact of recommender systems on user preferences was studied by Adomavicius et al. (2013). It stretches credulity to say that recommender system designers do not know about their nudging power. The survey of nudging mechanisms in recommender systems by Jesse and Jannach (2020) shows just that.

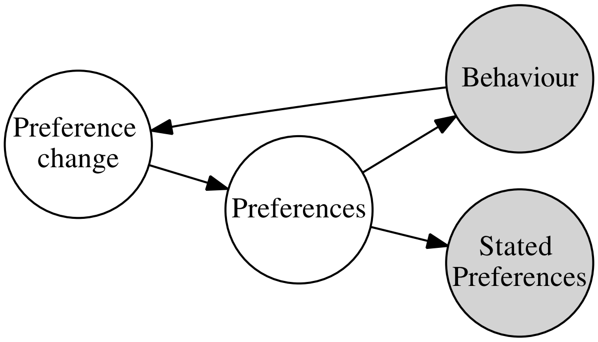

Figure 1: A Causal diagram showing the relationship between Preferences, Preference Change, Stated Preferences and Behaviour. Only Variables in grey can be observed. Preferences influence behaviour, but behaviour can cause preference change.

Public discomfort concerning the practice of private companies manipulating user behaviour is beginning to be reflected in regulation. Article 5 of the EU draft AI Act 2021 prohibits the use of an AI system that *deploys subliminal techniques beyond a person’s consciousness in order to materially distort a person’s behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.* At present, uncertainties exist about almost every aspect of this provision and how it will be enforced.

Given that behaviour change is possible, behaviour can influence preferences and preferences change anyway in response to exogenous events, it seems strange that models of preference change are few and far between. Jacobs (2016) provides one of the few dedicated literature reviews on the subject that we could find. Perhaps this is because empirical evidence concerning the effect of deployed AI systems is hard to find. This is puzzling given the generally acknowledged explosion in data collection possibilities that modern technology has enabled. Kramer, Guillory, and Hancock (2014) demonstrated that users' moods could be manipulated by changing what appeared on their Facebook news feed. The ensuing public and academic reception to the deliberate altering of people's moods without telling them was understandably not positive (Verma 2014). Consequently direct sources of proprietary data concerning the effect of Algorithm design on user preferences have not been forthcoming for public research. Other obstacles exist; the US Supreme Court recently ruled in *Van Buren v. United States* (2021) that certain academic research on web platforms would be protected from prosecution under Computer Fraud and Abuse Act 1986 (CFAA) (Villasenor 2021). Researchers can now devise programs to monitor user-facing algorithms without fear of custodial jail sentences but are still not party to the large scale behavioural data which would shed light on behavioural and by extension preference change.

One could argue that since the incentives of governments and large companies are not aligned with those of their users, behaviour and preference change externalities are inevitable. We will later argue that even a developer of an AI system whose only objective is to learn the preferences of their users is just as prone to manipulating their users' preferences as someone who is targeting behaviour or preference change for profit. Firstly we will consider in more detail the mechanisms that alter human preferences.

The mechanisms that manipulate preference

In this section we will briefly identify the most likely mechanisms which alter user preferences predominantly in the simple case of content-recommenders. We posit that preference manipulation comes from two separate sources which combine efficiently: 1) the mechanics of the recommender algorithm itself and 2) the generator of the content. There is a symbiosis between content generators who generate popular content and recommender systems that can alter preferences to fit that content. For the most part recommender system owners do not yet create content though there are some exceptions. Netflix amongst other video content platforms will use its analytics to make more addictive shows. Some internet retailers may choose to design and retail their own branded goods using their privileged data and product placement powers. In the near future, the advent of improved generative text and video technology can drastically lower the cost of developing and prototyping content and facilitate the exploration of novel manipulation techniques for media content platforms in an end to end automated manner.

One feature of recommender systems independent of preference plasticity is the phenomenon of popularity bias, whereby certain popular items are recommended more often than less popular items. This allows popular items to grow ever more popular (Abdollahpouri, Burke, and Mobasher 2017; Mansoury et al. 2020) and the process reinforces itself. This is a symptom of a wider problem with recommender systems - confounded data. The behaviour data used to train and test algorithms has already been influenced by the algorithm; this creates an amplifying feedback loop which increases homogeneity of recommended content (Chaney, Stewart, and Engelhardt 2018). In summary, naive recommenders have a natural tendency to push people towards the same small set of content and user's experiences are homogenised (Abdollahpouri 2019).

The mere-exposure effect describes the tendency for people to adapt preferences towards things they are familiar with (Fang, Singh, and Ahluwalia 2007). Related and similar effects are the availability bias (Tversky and Kahneman 1973), anchoring (Furnham and Boo 2011)) and the recognition heuristic (Goldstein and Gigerenzer 2002). This suggests that a content recommender that increases homogenisation for certain users would change the preferences of those users to whatever narrow band of content they are being recommended.

The combination of a recommender amplifying a few popular items and humans changing their preferences to the things which they are familiar with (ie recommended more often) is a powerful combination. However it does not explain the popularity of extreme content and the emergence of polarisation. Looking at the specific effects of content types, it seems certain types of content are more likely to lead to preference change than others. For example, it has been shown that conspiracy theory content is particularly potent (van der Linden 2015);van Prooijen and van Vugt (2018) hypothesise this predilection is for evolutionary reasons. Similarly content purporting to be from an impartial news source is effective at altering people's preferences; Alfano, Carter, and Cheong (2018) call this top-down technological seduction. Content which engenders strong emotion is likely to have manipulative effects on user preferences (Kusev et al. 2017). It is alleged that Facebook's newsfeed algorithm prioritised content that had received angry face emojis to maximise user engagement (Merill and Oremus 2021). So serious is the problem, Roozenbeek and van der Linden (2021) consider the effects of such content types as a matter of international security.

This is a simple account of preference change dynamics and ignores other mechanisms which tap into the many psychological biases and heuristics that humans have been shown to reliably exhibit. Alfano, Carter, and Cheong (2018) for example point to auto completion systems as ways of grouping users together and pushing them in certain directions. Other research have looked to the study of social effects where groups of people with similar views can coalesce leading to similar effects of polarization driven by confirmation bias (Del Vicario et al. 2017).

The discussion has so far been focused on recommender type dynamics where users are served content and their preferences are inferred through their observed behaviour. Preference elicitation is also vulnerable to behaviour manipulation techniques and has been more widely studied. Perhaps most famously people's numerical estimates can be adjusted based on prior exposure to higher or lower numbers using the anchoring effect (Furnham and Boo 2011). A simple example of this in practice is the suggested donation figures routinely used on donation forms. Perhaps most damning was the finding by (Hall, Johansson, and Strandberg 2012) that even after having given their preferences, when they were secretly changed by the experimenters, participants would often alter their views to match their (falsely recorded) ones. In short People can be told what their preferences are and they will change them.

None of the preference change mechanisms in this section are particularly complicated. In the cases of recommenders it amounts to repeating content types which claim to be true to users to the exclusion of other content types. We do not think that this scheme was intentional from the outset, it has just occurred. This begs the question, could an AI reproduce preference manipulation from scratch? We think a generative text algorithms would recover many human preference or behaviour manipulation techniques (framing for instance) with high regularity. Even a simple AI could just make up what it thought its users' preferences were and provided they believed it had listened to them in the first place, adjust whatever they did really think to the AI's choice. The question is why would an AI system be incentivised to intend to change human preferences?

Value Alignment and Preferences

An area where computer scientists are very interested in learning preferences is the subject of value alignment. The value alignment problem concerns the difficulty of writing objective functions for AI systems which prevent undesirable behaviour or allows AI to solve tasks that are otherwise hard to describe. In practice, as Lehman, Clune, and Misevic (2020) show, AI systems have a reliable habit of cheating to find solutions to given objectives.

We don't believe the observed problems surrounding recommender systems in the previous section are examples of the alignment problem. Though often unintended, the changes brought in user preferences are favourable for system owners, principally by making users more predictable. The algorithms are doing what they were designed to do - make money efficiently for their owners by increasing the time their users spend online. In common with many persistent externalities, the measurement and valuation of the harm caused is difficult. Russell, Dewey, and Tegmark (2015) calls this a *validity* problem; *"validity is concerned with undesirable behaviours that can arise despite a system's formal correctness"*.

One approach to the problem of value alignment is Inverse Reinforcement Learning (IRL); the construction of a human's utility function or values by the observation of their behaviour (Ng and Russell 2000). IRL is hard, it is an ill-posed problem in that any number of solutions (utility functions) can explain a given observed behaviour set in a single setting (Ng and Russell 2000) though it can be shown with a wide enough variety of settings, utility functions can be faithfully recovered (Amin and Singh 2016). Even so, certain assumptions need to be made about rationality, else as Armstrong and Mindermann (2019) show, any algorithm that derives a utility function could be arbitrarily bad at recovering an agent's actual utility. Hadfield-Menell et al. (2016) present Cooperative Inverse Reinforcement Learning as a better way of achieving alignment. (Russell 2020) presents three principles for AI developers to create beneficial machines which all rely on preferences:

1. *The machine's only objective is to maximise the realization of human preferences.*
2. *The machine is initially uncertain about what those preferences are.*
3. *The ultimate source of information about human preferences is human behavior.*

The difficulty in applying these principles is the causal relationship between behaviour and preferences as in Figure 1; behaviour indicates preferences but behaviour change begets preference change.

Given the non-stationarity and plasticity of human-preferences, any AI/ML approach to the learning of preferences seems to have a difficulty at its heart. Preference measurement takes time and the process might affect them, in other words, preference elicitation efforts suffer from the Observer Effect (Salkind 2010). This also includes any other techniques concerned with the elicitation and representation of preferences such as CP-nets (Boutilier et al. 2004; Loreggia et al. 2018) and active learning type efforts (Sadigh et al. 2017; Christiano et al. 2017). More problematically the AI/ML system is not often neutral to the preferences it learns, as Russell states: *"like any rational entity, the Algorithm learns how to modify the state of its environment - in this case the user's mind - in order to maximise its own reward"*. The same effect is noted in Soares (2016): "*Actions which manipulate the operator to make their preferences easier to fulfil may then be highly rated, as they lead to highly-rated outcomes (where the system achieves the operator's now-easy goals)*". Further back in time still Yudkowsky (2011) note that AI might rewire a programmers' brains to fulfil the objective of maximally pleasing them. Krueger, Maharaj, and Leike (2020) term this *Auto-Induced Distributional Shift*. This effect has been modelled by Everitt et al. (2021) using causal influence diagrams. With this technique, situations can be identified where there is a 'Instrumental Control Incentive' over user behaviour/preferences, that is to say settings where an algorithm has an incentive to alter the behaviour of the users it models in order to maximise its own objective function. Evans and Kasirzadeh (2021) show this to occur in the case of a recommender system trained through Reinforcement Learning. In a process that the authors term *user-tampering*, the recommender polarises its users in order to increase their predictability. This is also shown to be the case by Jiang et al. (2019) with a multi arm bandit learning model.

We make the observation that in practice AI/ML systems are often inter-temporal in their nature regardless of whether the learning algorithm behind them explicitly recognises multiple periods or not. Users will reuse a system over time and therefore their preferences will change as they adapt to the system. Commercial systems are typically iterated in practice, with a constant program of minor design improvements, A/B testing and retraining. Unless a particular effort is made to measure a users preferences before they begin interacting with an AI/ML system, it becomes impossible to know whether the system is doing a really good job or whether the system has just altered the preferences of its users to do a really good job.

Which preferences should be learned when the topic of preferences learning arises? The preferences that might exist before the user came into contact with the preference elicitation system or preferences after they have been altered? The instinctive response is to say the former, and that is suggested as solution by Everitt et al. (2021) to the problem of altering user preferences/behaviour to suit an objective. An alternative is the impact regularizer of Amodei et al. (2016) or a low impact learner (Armstrong and Levinstein 2017) which would seek to minimise the effect of the system on preferences. Neither solution is perfect because they might deny the legitimacy of a user's changed preferences. To use the example of a video recommender system, it could be the case that a user learns something after watching something and their preferences change as a result. Serving content to them as if they couldn't change could be just as bad as serving them content that targets prolonged engagement since it might trap users in a certain category of content.

Efforts are beginning to be made to address these problems. On the subject of non-stationary preferences, Chan et al. (2019) present a bandit algorithm to aid in the situation where a user is unsure about their preferences. In chapter 9, Russell (2020) discusses the problems associated with preference change and the difficulty with assigning moral valence to it. Perhaps here the more visible discussion surrounding the ethics of behaviour change can help. Resources are available to assess what constitutes good and bad behaviour change (Lades and Delaney 2020). The problem has been also considered for a long time in Welfare Economics and the philosophy of autonomy through the prism of Adaptive Preference Formation which was originally a rejection of utilitarianism (Teschl and Comim 2005). Elster (2016) develops a theory to separate more desirable preference changes like those caused by learning and experience from some of the less desirable ones that this article has touched on. Colburn (2011) characterises adapted preferences as those formed through covert influence and therefore undermine autonomy because users have not consciously chosen them. As Russell puts it, for an AI to learn preferences safely, it must be given some preferences over the type of preference changes that are allowed. For this to occur, the causes of any preference change need to be understood.

Conclusion

This article makes the observation that AI/ML practitioners often make an implicit assumption that preferences are static artefacts which can be learned with no effect on them. Sometimes preferences are learned from stated preference data, but more often than not they are learned from data concerning behaviour - i.e., assuming some variant of Revealed Preference Theory and rational behaviour assumptions. The assumption of non-changeable preferences is at odds with the behavioural change complex whose founding principle is that user behaviour can be manipulated. Systems that learn user preferences are at best likely to impact them during the process and at worst are likely to manipulate them to suit their own objective function in a process called Auto-induced Distributional Shift. Without an effort to record user preferences over time, it is difficult to know whether a particular system is doing its task well or altering user preferences to make its task easier.

A more considered approach to preference change in computer science is emerging, born from concerns surrounding Artificial General Intelligence (AGI) and value alignment. These are well founded since we have seen how user manipulation has already been effected by very limited algorithms. Theoretical and empirical research concerning the impact of recommender systems does recognise preference/behaviour change as a cause of problems like user polarisation. Companies are not incentivised to share data on such a sensitive topic, so much of the research on the topic has necessarily required multi agent simulations. This type of research is not without its critics due to its non-standardised approach (Winecoff et al. 2021) and has open challenges (Chaney 2021). We believe the validity of the results produced by simulations depend on the realism of their user preference change mechanisms. As a priority, a cross-disciplinary effort grounded on Empirical research is required to understand these processes as proposed by Franklin et al. (2022).

References

Abdollahpouri, H. 2019. Popularity Bias in Ranking and Recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 529–530. Honolulu HI USA: ACM.

Abdollahpouri, H.; Burke, R.; and Mobasher, B. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 42–46. Como Italy: ACM.

Adomavicius, G.; Bockstedt, J. C.; Curley, S. P.; and Zhang, J. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, 24(4): 956–975.

Albarracín, D.; and Jr, R. S. W. 2000. The Cognitive Impact of Past Behavior: Influences on Beliefs, Attitudes, and Future Behavioral Decisions. *Journal of Personality and Social Psychology*, 79(1): 5–22.

Albarracín, D.; and McNatt, P. S. 2005. Maintenance and Decay of Past Behavior Influences: Anchoring Attitudes on Beliefs Following Inconsistent Actions. *Personality and Social Psychology Bulletin*, 31(6): 719–733.

Alfano, M.; Carter, J. A.; and Cheong, M. 2018. Technological Seduction and Self-Radicalization. *Journal of the American Philosophical Association*, 4(3): 298–322.

Alfano, M.; Fard, A. E.; Carter, J. A.; Clutton, P.; and Klein,

C. 2020. Technologically scaffolded atypical cognition: the case of YouTube's recommender system. *Synthese*.

Amin, K.; and Singh, S. 2016. Towards Resolving Unidentifiability in Inverse Reinforcement Learning. *arXiv:1601.06569 [cs]*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]*.

Ariely, D.; and Norton, M. I. 2008. How actions create – not just reveal – preferences. *Trends in Cognitive Sciences*, 12(1): 13–16.

Armstrong, S.; and Levinstein, B. 2017. Low Impact Artificial Intelligences. *arXiv:1705.10720 [cs]*.

Armstrong, S.; and Mindermann, S. 2019. Occam's razor is insufficient to infer the preferences of irrational agents.

Atkins, L.; Francis, J.; Islam, R.; O'Connor, D.; Patey, A.; Ivers, N.; Foy, R.; Duncan, E. M.; Colquhoun, H.; Grimshaw, J. M.; Lawton, R.; and Michie, S. 2017. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. *Implementation Science*, 12(1): 77.

Beshears, J.; Choi, J., J.; Laibson, D.; and Madrian, B. C. 2008. How are preferences revealed? *NBER Working Paper Series*.

Bleidorn, W.; Hopwood, C. J.; and Lucas, R. E. 2018. Life Events and Personality Trait Change: Life Events and Trait Change. *Journal of Personality*, 86(1): 83–96.

Boutilier, C.; Brafman, R. I.; Domshlak, C.; Hoos, H. H.; and Poole, D. 2004. CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements. *Journal of Artificial Intelligence Research*, 21: 135–191.

Caplin, A.; Dean, M.; and Martin, D. 2011. Search and Satisficing. *American Economic Review*, 101(7): 2899–2922.

Chan, L.; Hadfield-Menell, D.; Srinivasa, S.; and Dragan, A. 2019. The Assistive Multi-Armed Bandit. *arXiv:1901.08654 [cs, stat]*.

Chaney, A. J. B. 2021. Recommendation System Simulations: A Discussion of Two Key Challenges. *arXiv:2109.02475 [cs]*.

Chaney, A. J. B.; Stewart, B. M.; and Engelhardt, B. E. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *Proceedings of the 12th ACM Conference on Recommender Systems*, 224–232.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *arXiv:1706.03741 [cs, stat]*.

Cialdini, R. B.; and Trost, M. R. 1998. Social influence: Social norms, conformity and compliance. In *The handbook of social psychology*. Mcgraw-Hill.

Colburn, B. 2011. Autonomy and Adaptive Preferences. *Utilitas*, 23(1): 52–71.

Del Vicario, M.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2017. Modeling confirmation bias and polarization. *Scientific Reports*, 7(December 2016): 1–9.

Dolan, P.; and Galizzi, M. M. 2015. Like ripples on a pond: Behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47: 1–16.

Elster, J. 2016. *Sour Grapes: Studies in the subversion of rationality*. Cambridge: Cambridge University Press.

Evans, C.; and Kasirzadeh, A. 2021. User Tampering in Reinforcement Learning Recommender Systems. *arXiv:2109.04083 [cs]*.

Everitt, T.; Carey, R.; Langlois, E.; Ortega, P. A.; and Legg, S. 2021. Agent Incentives: A Causal Perspective. In *AAAI Conference on Artifical Intelligence*.

Fang, X.; Singh, S.; and Ahluwalia, R. 2007. An Examination of Different Explanations for the Mere Exposure Effect. *Journal of Consumer Research*, 34(1): 97–103.

Franklin, M.; Ashton, H.; Gorman, R.; and Armstrong, S. 2022. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. *AAAI-22 Workshop on AI For Behavior Change*.

Furnham, A.; and Boo, H. C. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1): 35–42.

Goldstein, D. G.; and Gigerenzer, G. 2002. Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1): 75–90.

Gui, M.; Shanahan, J.; and Tsay-Vogel, M. 2021. Theorizing inconsistent media selection in the digital environment. *The Information Society*, 37(4): 247–261.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29: 3909–3917.

Hall, L.; Johansson, P.; and Strandberg, T. 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE*, 7(9).

Hill, T.; Kusev, P.; and van Schaik, P. 2019. Choice Under Risk: How Occupation Influences Preferences. *Frontiers in Psychology*, 10: 2003.

Jacobs, M. 2016. Accounting for Changing Tastes: Approaches to Explaining Unstable Individual Preferences. *Review of Economics*, 67(2): 121–183.

Jesse, M. W.; and Jannach, D. 2020. Digital nudging with recommender systems: Survey and future directions. *arXiv*, 1–28.

Jiang, R.; Chiappa, S.; Lattimore, T.; György, A.; and Kohli, P. 2019. Degenerate feedback loops in recommender systems. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 383–390.

Kozyreva, A.; Lewandowsky, S.; and Hertwig, R. 2020. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, 21(3): 103–156.

Kramer, A. D. I.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29): 8788–8790.

Krueger, D.; Maharaj, T.; and Leike, J. 2020. Hidden Incentives for Auto-Induced Distributional Shift. *arXiv:2009.09153 [cs, stat]*.

Kusev, P.; Purser, H.; Heilman, R.; Cooke, A. J.; Van Schaik, P.; Baranova, V.; Martin, R.; and Ayton, P. 2017. Understanding Risky Behavior: The Influence of Cognitive, Emotional and Hormonal Factors on Decision-Making under Risk. *Frontiers in Psychology*, 8.

Lades, L. K.; and Delaney, L. 2020. Nudge FORGOOD. *Behavioural Public Policy*, 1–20.

Lehman, J.; Clune, J.; and Misevic, D. 2020. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2): 274–306.

Lewis, P.; and McCormick, E. 2018. How an ex-YouTube insider investigated its secret algorithm. *Guardian*.

Loreggia, A.; Mattei, N.; Rossi, F.; and Venabe, K. B. 2018. A Notion of Distance Between CP-nets. In *Proceedings of AAMAS*, 7.

Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; and Burke, R. 2020. Feedback Loop and Bias Amplification in Recommender Systems. *arXiv:2007.13019 [cs]*.

Mathur, A.; Moschis, G. P.; and Lee, E. 2003. Life events and brand preference changes. *Journal of Consumer Behaviour*, 3(2): 129–141.

Merill, J. B.; and Oremus, W. 2021. Five points for anger, one for 'like': How Facebook's formula fostered rage and misinformation. *The Washington Post*.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. *ICML*, 1.

Nishimura, H. 2018. The transitive core: Inference of welfare from nontransitive preference relations: The transitive core. *Theoretical Economics*, 13(2): 579–606.

OECD. 2017. *Behavioural Insights and Public Policy: Lessons from Around the World*. OECD.

Orlowski, J.; Coombe, D.; and Curtis, V. 2020. The Social Dilemma.

Roozenbeek, J.; and van der Linden, S. 2021. Inoculation Theory and Misinformation. Technical report, NATO Strategic Communications Centre of Excellence.

Ruggeri, K. 2018. *Behavioral insights for public policy: concepts and cases*. Routledge.

Russell, S. 2020. *Human Compatible*. Penguin, 1st edition.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*, 36(4): 105–114.

Sadigh, D.; Dragan, A.; Sastry, S.; and Seshia, S. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation.

Salkind, N. 2010. *Encyclopedia of Research Design*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.

Schneider, C.; Weinmann, M.; and vom Brocke, J. 2018. Digital nudging: guiding online user choices through interface design. *Communications of the ACM*, 61(7): 67–73.

Soares, N. 2016. The Value Learning Problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016)*.

Sunstein, C. R. 2014. *Why Nudge: The Politics of Libertarian Paternalism*. Yale University Press.

Sunstein, C. R. 2016. *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press.

Sutherland, R. 2019. *Alchemy: The Surprising Power of Ideas that Don't Make Sense*. Random House.

Teschl, M.; and Comim, F. 2005. Adaptive Preferences and Capabilities: Some Preliminary Conceptual Explorations. *Review of Social Economy*, 63(2): 229–247.

Thaler, R. H.; and Sunstein, C. R. 2003. Libertarian Paternalism. *The American Economic Review*, 93(2): 175–179.

Thaler, R. H.; and Sunstein, C. S. 2008. *Nudge*. Yale University Press.

Tversky, A.; and Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2): 207–232.

Tversky, A.; and Kahneman, D. 1985. The Framing of Decisions and the Psychology of Choice. In Wright, G., ed., *Behavioral Decision Making*, 25–41. Boston, MA: Springer US.

van der Linden, S. 2015. The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases prosocial behavior and science acceptance. *Personality and Individual Differences*, 87: 171–173.

van Prooijen, J.-W.; and van Vugt, M. 2018. Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science*, 13(6): 770–788.

Varian, H. 2006. Revealed Preference. In *Samuelsonian economics and the twenty first century*, 99–115. Oxford University Press.

Verma, I. M. 2014. Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29): 10779.

Villasenor, J. 2021. Reining in overly broad interpretations of the Computer Fraud and Abuse Act.

Winecoff, A. A.; Sun, M.; Lucherini, E.; and Narayanan, A. 2021. Simulation as Experiment: An Empirical Critique of Simulation Research on Recommender Systems. *arXiv:2107.14333 [cs]*.

Wyer, R. S.; Xu, A. J.; and Shen, H. 2012. The Effects of Past Behavior on Future Goal-Directed Activity. In *Advances in Experimental Social Psychology*, volume 46, 237–283. Elsevier.

Yudkowsky, E. 2011. Complex Value Systems in Friendly AI. In Schmidhuber, J.; Thórisson, K. R.; and Looks, M., eds., *Artificial General Intelligence*, volume 6830, 388–393. Berlin, Heidelberg: Springer Berlin Heidelberg.