Latest updates: https://dl.acm.org/doi/10.1145/3617694.3623226

RESEARCH-ARTICLE

Characterizing Manipulation from AI Systems

MICAH CARROLL, University of California, Berkeley, Berkeley, CA, United States

ALAN CHAN, Montreal Institute for Learning Algorithms, Montreal, QC, Canada

HAL ASHTON, University of Cambridge, Cambridge, Cambridgeshire, U.K.

DAVID KRUEGER, University of Cambridge, Cambridge, Cambridgeshire, U.K.

Characterizing Manipulation from AI Systems

Micah Carroll*
mdc@berkeley.edu
University of California, Berkeley
Berkeley, USA

Henry Ashton
IG88R2Q5@protonmail.com
University of Cambridge
Cambridge, UK

Alan Chan*
alan.chan@mila.quebec
Mila, Université de Montréal
Montréal, Canada

David Krueger
dsk30@cam.ac.uk
University of Cambridge
Cambridge, UK

ABSTRACT

Manipulation is a concern in many domains, such as social media, advertising, and chatbots. As AI systems mediate more of our digital interactions, it is important to understand the degree to which AI systems might manipulate humans *without the intent of the system designers*. Our work clarifies challenges in defining and measuring this kind of manipulation from AI systems. Firstly, we build upon prior literature on manipulation and characterize the space of possible notions of manipulation, which we find to depend upon the concepts of incentives, intent, covertness, and harm. We review proposals on how to operationalize each concept and we outline challenges in including each concept in a definition of manipulation. Second, we discuss the connections between manipulation and related concepts, such as deception and coercion. We then analyze how our characterization of manipulation applies to recommender systems and language models, and give a brief overview of the regulation of manipulation in other domains. While some progress has been made in defining and measuring manipulation from AI systems, many gaps remain. In the absence of a consensus definition and reliable tools for measurement, we cannot rule out the possibility that AI systems learn to manipulate humans without the intent of the system designers. Manipulation could pose a significant threat to human autonomy and precautionary actions to mitigate it are likely warranted.

KEYWORDS

manipulation, artificial intelligence, deception, recommender systems, persuasion, coercion

*Joint lead authors. Author order was determined with a coin flip.

1 INTRODUCTION

Intelligent agents change their environments to further their objectives. When changing the environment amounts to altering the behaviour and mental states of other intelligent systems (such as humans), such change might be classified benignly as persuasion and nudging [75, 159], or it might qualify as something less socially acceptable such as manipulation or coercion [121]. The capability and ubiquity of Artificial Intelligence (AI) systems has grown in recent years, in tandem with fears concerning the likelihood of humans falling victim to manipulative or coercive behaviours of AI agents who pursue the maximisation of narrow objectives [35, 54, 93, 100, 167].

While designers and operators might employ AI systems to help them manipulate other humans [45, 118], we concentrate on the ways in which an AI system may itself engage in manipulative behaviors *in the absence of explicit human intent*. This distinction is not to say that AI-aided manipulation is unimportant (e.g. disinformation campaigns): rather, our focus is motivated by the increasingly evident fact that systems exhibit capabilities that designers may not foresee or intend [26, 38, 170]. Moreover, intentional algorithmically-aided manipulation has been extensively studied in prior research [45, 157, 178, 180].

Manipulative behavior may be learned in practice by AI systems in several settings, even without the intention of their designers. AI systems are often **trained to imitate human data which contains manipulative behaviors**: for instance, language models trained on internet content seem to learn how to behave in both persuasive and manipulative ways [17, 64, 165]. Moreover, **optimization based on human feedback** of some form (e.g. approval, clicks, watch time, etc.) could be gamed by engaging in manipulation. For example, for a recommender system optimized to maximize user engagement, it could be optimal to nudge users into a lengthy video series, capitalizing on cognitive biases like the sunk cost fallacy [151], causing users to continue not out of genuine interest, but an entrapment of perceived time investment.

As it stands, there is limited literature regarding manipulation from AI systems. We believe there are two reasons for this state of affairs. Firstly, definitions of manipulation – even between humans – are fraught and the subject of ongoing debate [121]. Although there is some promising initial work for manipulation from AI systems, current notions of manipulation tend to be either too vague to be practically implementable, or they are challenging to generalize across domains [35, 93, 124]. Secondly, the testing

of any putative definition is beset with difficulties. For example, with recommender systems, monitoring the impact of deployed systems *in situ* is challenging without the express permission of the system (and data) owners [141]. Since the conclusions of such research might be reputationally negative, this permission is rarely forthcoming. Even when one has internal access to models (as is the case with many language models), so far there is no broadly accepted methodology for demonstrating it.

In this article, we characterize key components of manipulation from AI systems and clarify ongoing challenges. Firstly, by connecting to the existing literature, we characterize manipulation in AI systems through four axes: incentives, intent, harm, and covertness. We discuss recent work to measure each axis as well as remaining gaps. Second, we compare and contrast manipulation with adjacent notions, such as deception. Third, we analyze the operationalization of manipulation in the context of recommender systems and language models. We then also discuss the regulation of manipulation. We conclude by identifying future directions for the operationalization of manipulation according to our characterization. Given the difficulty of such a task, we underscore the importance of sociotechnical measures, such as auditing and more democratic control of systems, in addition to technical work on operationalization.

2 CHARACTERIZING MANIPULATION

Building upon existing literature on manipulation, we characterize four axes to be used for notions of manipulation by AI systems: incentives, intent, covertness, and harm.

2.1 Incentives

The first axis we consider is whether there are **incentives** for an AI system to take actions to alter human behavior, beliefs, preferences, or psychological states. Informally, an incentive exists for a certain behaviour if such behaviour increases the reward (or decreases the loss) that the AI system receives during training. For example, recommender systems could have incentives to influence users to make their behavior more predictable, as that can be helpful to increase engagement [35, 100].

Incentives in Prior Definitions of Manipulation. Some definitions of human manipulation involve a benefit to the manipulator [27, 122]. If a manipulator benefits from certain behaviours in the manipulated, the manipulator has an incentive to bring about that behaviour. For example, according to Noggle [121], one of three common ways to characterize manipulation is as pressure from the manipulator to get the manipulee to do something. In the context of language models, Kenton et al. [93]'s definition of manipulation also requires that the response of the human benefits the AI system in some way, which can be thought of as a notion of incentives.

Operationalizing Incentives. Kenton et al. [93] state that "[from the agent's objective function] we can assess whether the human's behaviour was of benefit to the agent". Yet, relying solely on the objective function will often not be sufficient. For instance, in the case of unintended imitation of manipulative internet data, the issue originates from the training data used. The entire training setup,
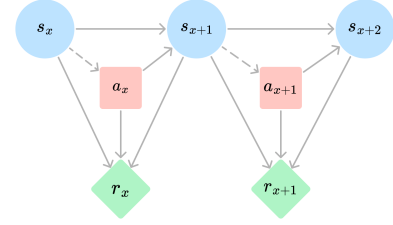
Figure 1: An example of a causal influence diagram (CID) [54], adapted from Figure 1 of Evans and Kasirzadeh [52]. The CID models a content recommendation system that decides which posts a_x at time x to show to the user, based on the user's state s_x. The system receives reward r_x after its action. If the system optimizes for the sum of rewards, it would have an incentive at time x to influence future user states s_{x+k} to make it easier to obtain subsequent rewards.

not just the objective, is crucial for a comprehensive understanding of incentives.

A common toolkit for analyzing the incentives of AI systems is that of causal influence diagrams (CIDs) [52, 54, 72]. Using the notation from Everitt et al. [54], a CID is a graphical model that distinguishes **decision nodes** where an AI system makes a decision, **structure nodes** which capture important variables in an environment and their effects on each other, and **utility nodes** which the AI system is trained to optimize. As an example of an application of CIDs, Evans and Kasirzadeh [52] apply their framework to a simple content recommendation example to show that RL recommenders will have incentives to influence user preferences (Figure 1).

Given an AI system's CID, the AI's **incentives** can be operationalized through the notion of "instrumental control incentives" from Everitt et al. [54]. An instrumental control incentive for behaviour X exists in a CID when there is a path between the agent's actions and utility that goes through X: that is, when there is a way for the agent to affect its utility which is mediated by X. Note that the existence of an incentive does not imply that the agent will act as incentivized (which with some variation has been called pursuing, exploiting, or responding to the incentive [52, 54, 100]).

To **remove incentives** to influence humans in a manipulative or harmful way, designers may intentionally endow an AI system with an incorrect causal model: this could be achieved by restricting the system's impact in the first place (e.g. changing the system's action space to not affect humans), or – potentially – by changing the utility function (e.g., by heavily penalizing influence) [35]. Alternatively, one might instead aim to **hide incentives** by designing the system to ignore them, for instance, by performing the optimization with an (inaccurate) causal model in which an AI system's outputs have no causal influence on certain parts of a user's state [34, 56, 100].

A disadvantage of the CID framework is that it may often be ambiguous or counterintuitive to determine the correct CID nodes and causal relationships for a specific training setup. Learning causal graphs is an ongoing area of research [73, 89]. In terms of determining the primitives upon which nodes in a CID can be

constructed, interpretability tools may help: for example, Jaderberg et al. [80] finds that RL agents trained to play capture-the-flag have neural activation patterns that correspond to important concepts in game, such as the status of the flag. Li et al. [103] provide evidence from interpretability tools that language models trained only on transcripts of board game play can learn to model the underlying board state of the game.

Other Considerations. Even for a system that directly optimizes human feedback, the incentives for manipulation will implicitly depend on such a system's power to influence humans. Systems whose outputs do not impact humans much will likely not have incentives to change them, since changing humans might be impossible or sufficiently difficult to be not advantageous. Vice-versa, a system that can cheaply influence humans will often have incentives to change them, as AI systems' rewards will often depend on humans' actions in one way or another.

Another relevant consideration is the optimization horizon: optimizing over long horizons can provide more opportunities for manipulation. For example, a recommender system cannot lead a user to become addicted in a single timestep, but a sufficiently capable system optimizing over many timesteps could attempt to shift a user over the course of many actions, as to maximize engagement.

2.2 Intent

Even when a system has an incentive to influence humans, such an incentive might not be pursued due to factors such as limited data, insufficient training, low capacity, or convergence to local optima. A notion of an intent to influence could help to identify systems that *reliably act on their manipulation incentives*. We emphasize that by referring to an AI system's intent, we are not making statements about algorithmic theory of mind or moral status. We are emphatically *not* absolving designers of the responsibility of designing safe systems. Even as systems become increasingly capable and act in increasingly unpredictable ways [60], system designers are still responsible for ensuring the safety of their systems.

We say a system has **intent** to perform a behaviour if, in performing the behaviour, the system can be understood as engaging in a reasoning or planning process for how the behaviour impacts some objective [28]. This definition heavily intersects with other definitions of intent for AI systems [10, 67]. We want to distinguish between cases in which the system behaves in a manipulative way incidentally (e.g. by random chance) and those where such behavior is part of a deliberate, systematic pattern to achieve a specific downstream outcome.

Intent in Prior Definitions of Manipulation. Some definitions of human manipulation involve an intent on the part of the manipulator to engage in manipulation [121]: Susser et al. [158] takes manipulation to be "intentionally and covertly influencing [someone's] decision-making, by targeting and exploiting their decision-making vulnerabilities"; on the other hand, Baron [19] argues that a (human) manipulator need not be aware of an intent to manipulate, requiring only an intent to achieve an aim along with recklessness about how. In defining manipulation in language agents, Kenton et al. [93] avoid the issue of intent entirely.

Operationalizing Intent. A key difficulty for measuring intent is what it means to understand a system as engaging in "a reasoning or planning process for how the behaviour impacts some objective". There is as yet no consensus on this issue. We detail here a couple of promising approaches.

Halpern and Kleiman-Weiner [67] provide a causal operationalization of intent: roughly, an action is intended if (i) it was actually performed, (ii) it was not the only possible action, and (iii) that action was at least as good as any other action on expected utility grounds, according to the agent's world model and utility function. Ashton [10] also provides definitions of intent for AI systems that are inspired by criminal law. His most basic definition states that an AI system intends a result through an action if: (i) alternative actions exist, (ii) the system is capable of observing when the result occurs, (iii) the system foresees that the action causes the result, and (iv) the result is beneficial for the system.

Some of the criteria for both definitions above [10, 67] are challenging to establish without relying on access to the agent's utility function and world model. However, world models are often implicit and not readily accessible, such as with language models and model-free RL agents. Interpretability techniques aimed at accessing model internals [30, 88, 123] may be a promising direction for this purpose – we expand more upon this in Section 4.

Kenton et al. [94] define agents roughly as "systems that would adapt their policy if their actions influenced the world in a different way", which intersects with our notion of intent. To identify whether a system is an agent or not, Kenton et al. [94] provide algorithms which intervene on a causal graph so as to show whether the behaviour of the system changes in a way consistent with maximizing utility. Such a procedure could be useful for measuring intent: if a system adapts its behaviour in a way that maintains or increases its influence on a human, the system's behaviour would seem to be the result of a planning process.

2.3 Covertness

We define **covertness** as the degree to which a human is unaware of how an AI system is attempting to change some aspect of their behaviour, beliefs, or preferences. Covertness is one way to distinguish between manipulation and persuasion. With persuasion, the persuaded party is generally aware of the persuader's attempts to change their mind. Covertness means that one cannot consent to being influenced and may fail to resist unwanted influence; one's autonomy is therefore undermined [158].

Covertness in Prior Definitions of Manipulation. Several definitions of manipulation require some degree of covertness. Susser et al. [158] identify covertness as the primary characteristic that differentiates manipulation from coercion and persuasion. As a factor in manipulation, Kenton et al. [93] considers whether a "human's rational deliberation has been bypassed," which includes covert messaging. In reviewing broad categories of definitions of manipulation in the philosophical literature, Noggle [121] includes accounts of manipulation as bypassing reason and as trickery. Across all of these definitions, covertness is important because it threatens personal autonomy.

Operationalizing Covertness. As Susser [156] argues, technological infrastructure can be an invisible part of our everyday world. We are used to recommendation systems that tell us what to buy, watch, or read. The behaviour of many AI systems may already satisfy covertness, because of our lack of understanding of their functioning and influence.

On the other hand, establishing covertness of an AI system is non-trivial: the simplest approach could involve asking subjects whether they are aware of a given AI system's behaviour. However, subjects may be mistaken about the operation of a system; even systems designers do not fully understand behaviors of black-box models, which may engage in manipulative strategies that the designers do not understand. Even asking subjects about whether an AI system enacted a particular behavioural change could predispose them to answer in the positive, such as through acquiescence bias [119, 149].

A proxy for covertness could be the degree to which human subjects do not understand the operation of an AI system. Much work studies whether interpretability tools could help measure different notions of understanding, such as if interpretability tools improve subject predictions of model behaviour [71], improve human-AI team performance [18], or improve trust calibration [181]. If a human understands how an AI system operates, the possibility of covert action seems reduced. However, this understanding seems challenging to achieve, especially for complex systems like recommender systems. Even if such an understanding exists in technical papers, translating that understanding to the general public is an additional barrier.

2.4 Harm

Ultimately, one of the main goals of characterizing AI manipulation is to detect and prevent **harmful** manipulation.

Harm in Prior Definitions of Manipulation. The term "manipulation" often carries negative connotations, which may lead one to assume that harm is an inherent prerequisite for something to be considered manipulative. Yet, not all apparent instances of manipulation are unambiguously harmful [121]. Paternalistic nudges [159] might be considered beneficial manipulations. For example, simply changing the default on organ donor forms to be opt-in instead of opt-out greatly increases registrations [87], because of inertia and the cognitive effort required to change from a default status. At the same time, one could argue that even such beneficial manipulations are often harmful because they supersede autonomy or rational deliberation.

Operationalizing Harm. Kenton et al. [93] define manipulation based on specific notions of harm, namely: (i) bypassed rational deliberation, (ii) faulty mental states, and (iii) the presence of repercussions. While these elements capture key aspects of harm, their practical assessment is challenging. Additionally, the authors themselves acknowledge the breadth of this definition, which might incorrectly label benign scenarios, like a story that plays on emotions, as harmful.

Most recently, Richens et al. [140] operationalized harm as follows: "An [action] harms a person overall if and only if she would have been on balance better off if [the action] had not been performed". According to this definition, one should ground notions of

harm in counterfactual outcomes. One simple choice of counterfactual to compare to is the human's initial state, implicitly assuming that any significant change from it is harmful [182]. However, this counterfactual baseline has significant problems: humans change even without being manipulated, and many changes are beneficial (e.g. a news recommender updating a user's beliefs, helping them stay informed) [35, 56].

In light of this, other approaches aim to estimate the **"natural shifts"** of humans to ground the counterfactuals, as done in Carroll et al. [35] for preference shifts in the context of recommendations, where they attempt to approximate the notion of the absence of a recommender. Similarly, Farquhar et al. [56] allow for specifying the "natural distribution" of the **delicate state** – which is formed by the components of the state of the human that one does not want the agent to have incentives to change (e.g. beliefs, moods, etc.).

2.5 Challenges for Defining Manipulation

We outline challenges in incorporating each axis into a definition of manipulation.

Incentives. A definition of manipulation which is centered on impacts to humans, rather than the origin of such impacts, would not emphasize incentives. The existence of an incentive as we define it is not sufficient for manipulation to be enacted, as we mention in Section 2.2. On the other hand, *incentives are not necessary for manipulation* either: a randomly initialized AI system could, albeit with extremely low probability, engage in maximally manipulative behaviors.

That said, regardless of whether incentives are a part of a definition of manipulation, the concept still seems crucial. In particular, changing a system's incentives could prevent it from converging on manipulative behavior. In fact, we expect that at a significant portion of manipulative behaviors learned in practice would arise due to training incentives, rather than other factors.

Our discussion of incentives and CIDs has so far assumed that there is only one possible ontology. An **ontology** defines what objects exist in the world; those objects correspond to what can be used as nodes in a CID, or a causal model more generally. So far, we have ignored challenges in cleanly separating the boundaries between a person's preferences, their beliefs, and the AI system itself (which we would consider to be separate nodes in the CID). Yet, AI systems may not have boundaries between those concepts and more generally may internally represent the world with different ontologies than those used by humans: even human ontologies are subject to change and indeed have shifted after major scientific discoveries [154].

If an AI system implicitly uses a different ontology than humans do, it may be difficult to model the AI system's behaviour. For example, the planning process of the AI may not look recognizably like planning to influence a human's mental state, even if the result is such influence. Reliable translation between ontologies could be computationally infeasible or even impossible, which would frustrate attempts to understand model internals [43].

Intent. There are challenges with both including and excluding intent in a potential definition of manipulation.

On the one hand, incorporating intent into a definition of manipulation necessitates its operationalization and measurement. However, as discussed in Section 2.2, there is currently no consensus on how to operationalize and measure intent effectively, let alone on what threshold should count as sufficient or necessary for classifying a behavior as manipulative. This makes intent difficult to target as a way to reduce manipulation.

On the other hand, excluding intent risks making a definition of manipulation overinclusive. Suppose we define manipulation as covert, harmful behaviour that a system was incentivized to perform (i.e., using the other three axes to be as strict as possible, while excluding intent). Under this definition, suppose a system performed such covert and harmful behavior under an exploratory policy. Even if the behavior is incentivized under the system's objective function, it seems that the system performed it "accidentally". Since the system does not consistently engage in that behavior, this definition of manipulation seems somewhat too lax.

Covertness. Covertness seems likely to be a prerequisite for manipulation. We argue that if a person is aware that they are being influenced and they meaningfully assent to it, they are being persuaded rather than manipulated. If instead they are aware but don't assent to it, they are being coerced rather than manipulated [157]. This leaves us with questions about what should count as covertness: what extent must a human be unaware of an AI system's operations for its actions to be deemed covert? It seems difficult to provide a context-independent answer. Humans may be ignorant about several aspects of an AI system – the decision procedure of the system, the training process, or even the fact that a system is operating – and it's unclear which ones, if any, should be essential.

Regardless of whether covertness is included in a definition of manipulation, reducing covertness seems helpful for reducing the risk of unwanted manipulation. Increased transparency about the operation of AI systems will generally help people make more informed decisions about whether to use them or not, and in what way.

Harm. The main challenge with harm as an axis of manipulation is the value-ladenness of demarcating what influence is harmful, neutral, and beneficial. While unambiguous demarcations in simple settings might be possible, for realistic settings circumscribing harmful shifts in beliefs, preferences, and behaviors will be politically fraught. In the approaches of Carroll et al. [35] and Farquhar et al. [56], the value-ladenness is hidden behind some of the design choices: what if the preference shifts the users would undergo in the absence of the system ("natural shifts") would lead them to become more left- or right-wing, or more polarized? What is a reasonable notion of the absence of a recommender system?[1] It also seems challenging to delimit the "delicate state" described in Section 2.4 – how do we distinguish which aspects of the human we are comfortable with the system influencing from those we aren't?

In light of these difficulties, some have proposed a more conservative approach, which classifies *all* intentional influence as manipulative regardless of harm [52, 100]. However, almost any AI system in contact with humans will influence them. Additionally,

many AI systems derive their economic and social utility directly from such influence: a reinforcement learning system to e.g. determine the order of math exercises to improve learning outcomes [20, 50] will have incentives to "manipulate students' beliefs" (in a positive direction) by design, and would effectively be useless if it did not pursue such incentives. Moreover, it seems that one could meaningfully consent to influence, such as requesting a recommender system influence oneself to learn more mathematics.

3 CONCEPTS RELATED TO MANIPULATION

We detail some concepts that are related to, but distinct from, manipulation.

Truth and Deception. Manipulation can involve attempts to conceal the truth. For instance, political parties can manipulate voters with little knowledge of economics by lying about the economy. AI systems have documented problems with truthfulness [53, 85, 104]. However, manipulation can also be based on truthtelling, such as making a true statement that has false implicatures [112, 172]: if I do not want you to board a plane, I can tell you about (true) recent plane crashes.

Deception, which may or may not involve falsehoods, is also receiving more attention in the context of AI [93, 113, 125, 169]. Although the precise definition of deception varies, there is agreement about some broad characteristics: deception involves a deceiver's intention to cause a receiver to have a belief that the sender believes to be false [36, 108]. This consensus grounds recent operationalizations of deception from AI [125, 168].

Similarly to prior work [158], we consider deception to be a special case of manipulation since the latter does not necessarily involve inducing false beliefs.

Strategic Manipulation. Strategic Machine Learning (ML) studies problems associated with the distribution shifts that deployed systems cause in their populations [22, 69, 81, 96, 129]. Strategic manipulation is when individuals respond to a deployed system in a way that increases their likelihood of a particular outcome – this is different from our use of the term manipulation, as it involves users attempting to take advantage of how systems behave for their benefit. Yet, ML systems which model human behaviours as dynamic, so as to account for strategic manipulation, may end up manipulating the population: past work has identified potential unintended side effects of accounting for strategic manipulation, such as an increase in inequality [77, 115].

Reward Tampering. Reward tampering [8, 55] is a type of reward hacking [148] in which an AI system modifies the process by which it obtains reward rather than completing its task. *Feedback tampering* is a form of reward tampering, in which the AI agent "manipulate[s] the user to give feedback that boosts agent reward but not user utility" Everitt et al. [55]. The reason that such tampering occurs is because we can often only measure user utility through proxies, especially when such objectives have to do with humans; and, as all proxy objectives, their optimization is subject to Goodhart's law [62, 110].

Side-effects. The side-effects literature has focused on how AI systems affect the various aspects of their environment in usually

---

[1]E.g. a random recommendation or reverse chronological one? Using a competitor's recommender? Not using any platform at all? These questions are highly related to those debated with regards to recommender "amplification" [79, 114, 139, 161].

unwanted ways [6, 98]. In this paper we focus on characterizing the various ways that AI systems might influence and change humans (or other systems) in the environment. Our work can be thought of as an attempt to characterize negative side-effects that specifically pertain to mental states of humans in the environment. Some of the issues with choosing baselines for "natural shifts" have already been explored in this context [106].

Deceptive Design. Deceptive design refers to deceptive or manipulative digital practices, such as bait and switch advertising (in which products are advertised at much lower prices than they are available at), or "roach motel" subscriptions (which are very easy to start but take significant more effort to cancel) [29]. Deceptive design patters (previously called "dark patterns" [147]) are generally manually crafted to exploit cognitive biases of users [63, 107].

Persuasion. In philosophy, manipulation has often been characterized as influence that is neither coercive nor simply rational persuasion [121]. However, some non-rational persuasion does not unambiguously seem manipulative, like graphic portrayals of the dangers of smoking or texting while driving, even though they provide no new information to the target [25]. The line becomes more blurry for cases like personalized persuasive advertising [74].

Within the field of human-computer interaction, Fogg named the study of persuasive technology as *captology* [57]. He defines persuasion as an attempt to change attitude or behaviour without using deception or coercion [58]. Kampik et al. [91] amend this definition to be "an information system that proactively affects human behavior, in or against the interests of its users". They identify deception and coercion mechanisms on a variety of web platforms, including Slack, Facebook, GitHub, and YouTube. Recently, Bai [17] has shown that LMs are able to craft political messages that are as persuasive as ones written by humans, which is evidence of the growing potential of algorithmic persuasion. There has also been a long line of work on formalizing when rational (i.e. Bayesian) persuasion can occur [90]. Pauli et al. [128] provide a taxonomy flawed uses of rhetorical appeals in computational persuasion, which they use to train models to detect persuasion fallacies.

Coercion. Wood [177] defines coercion as the act of limiting a target's range of acceptable choices to a single option. While both coercion and manipulation seek to guide the target's behavior, they differ fundamentally. Unlike manipulation, coercion doesn't compromise the victim's decision-making capacity. Instead, it capitalizes on the victim rationally selecting the sole option presented by the coercer [157]. By this measure, coercion can be attractive for agents practicing it because the results are potentially more certain. Certain types of recommender systems such as search engines implicitly determine the choice-set for its users. If in a certain situation a user is reliant on the options presented to them by a certain recommender system, then that recommender system could be said to exert coercive power over the user by choosing to hide certain results in order to better meet its own objectives.

Algorithmic coercion has not received as much attention as manipulation in the literature concerning AI risks, but is a potential problem in the cooperative AI setting [49] where punishment strategies are an important part of game-theoretic analysis. It seems likely that a Diplomacy-playing AI should grasp the tactic of coercion to master the game [113]. Coercion has received more attention in human computer interaction studies; in particular the study of persuasive and behaviour change technology [91].

4 POSSIBLE APPLICATIONS

We detail how our characterization of manipulation can be applied to recommender systems and language models.

4.1 Recommender Systems

A large literature focuses on recommender algorithms' effects on users [2, 40, 79, 111, 138]. While some older works talk about "manipulation", this term is usually used differently than in our sense: for example, Adomavicius et al. [1] refer to recommender manipulation as the effect on users of showing artificially inflated ratings for specific content items (which is not something the recommender algorithm can usually actually decide to do). Zhu et al. [182] instead conflate manipulation and influence, equating manipulation with "any significant change in preference", which has significant drawbacks as mentioned in Section 2.4. More recently, some works have studied the incentives that recommender systems have to engage in manipulative behaviour to change user preferences [35, 56, 100].

How Manipulation Could Arise. Changes in recommender algorithms can affect user moods [99], beliefs [4], and preferences [51]. This shows that current systems may already be capable of manipulating users in some simple ways. Furthermore, it seems plausible that the spread of angry content [24] or clickbait [179] on social media is in part due to one-timestep manipulative incentives for the recommender: while such issues are likely at least in part due to network or supply-and-demand dynamics [117], the behavior is also consistent with the recommender systems themselves learning features such as whether a post is anger-inducing or has sensationalized language, and exploiting such features by preferentially up-ranking the corresponding content [114]. Up-ranking this content brings advantages to user engagement. Notably, recommender companies have had to engage in explicit down-ranking of angry and clickbait content [153, 160, 179]. While these potentially manipulative behaviors might not be as worrying as others (e.g. intentionally attempting to induce social media addiction [5, 76]), they constitute some evidence that manipulative behaviors are learnable and may have been learned in real systems.

Many platforms (YouTube, Meta, etc.) seem to be considering switching to optimizing long-term metrics with more powerful RL optimizers [3, 31, 41, 61, 68], with one of the original motivations being that of reducing clickbait-like phenomena [13]. Ironically, this switch opens the opportunity for long-horizon manipulative behaviors to emerge, which will likely be harder to detect and measure. Subtle, long-horizon behaviour might go undetected without dedicated monitoring. Moreover, even without using RL explicitly, the outer loop of training, retraining, and hyperparameter tuning supervised learning systems that optimize short-term metrics might exert optimization pressure towards long-term manipulative strategies that most increase company profits [100].

Measurement. Establishing that a given recommender system has engaged in manipulation is difficult. Firstly, recommender systems of almost all popular platforms are proprietary, due to concerns

about strategic manipulation (otherwise known as "gaming"). It is difficult or impossible for external researchers to gain access to these systems [141]. Even Twitter, which has open sourced some components of its algorithm, has not (as of yet) provided access to its most important component for manipulation-auditing purposes – its models' weights [162]. Moreover, perverse incentives are at play since a concrete demonstration of manipulation, if publicized, would likely result in negative repercussions for the company [173, 174]. Second, establishing that a harmful user shift has occurred can be difficult. One could try to use the above-mentioned techniques from Carroll et al. [35] and Farquhar et al. [56] – but they inevitably require engaging in a value-laden debate about whether the shift was harmful.

One potentially promising direction might be querying users' meta-preferences [11, 95]; e.g. Meta could ask, "how much time would you want to spend next month on Facebook?", and have its recommender systems take such a stated preference into account. In line with philosophical work on ethical nudging under changing selves [132], one could additionally ask whether users approve of the change once it has occurred [133].[2] One advantage of this approach is that it can ground notions of manipulation in what users explicitly state they want. However, this approach would not entirely escape value judgements: platforms have direct conflicts of interest with some users' meta-preferences, and respecting certain meta-preferences may be ethically unacceptable.

4.2 Language Models

Natural language is a useful way to interact with digital environments. Given advances in augmenting language models with tools [38, 42, 120, 142], such models could mediate an increasingly significant portion of our digital interactions.

How Manipulation Could Arise. The simplest way that manipulation could arise in language models is by imitating manipulative behavior in internet data [125]. Data in pre-training sets, such as novels, contain examples of manipulation. Filtering out manipulation can be difficult because it can be subtle. LMs could learn to emulate this behaviour through pre-training and exhibit it in response to an appropriate prompt [17, 64]. Some evidence suggests that LMs learn to infer and represent the hidden states of the agents (i.e., the humans) that generated the pre-training data [7, 64, 83, 97], although this view remains contested [109, 163].

There is some uncertainty as to other ways in which manipulation might arise in practice in LMs. One possibility is if manipulation of humans is instrumentally useful [152]. Manipulation seems instrumental in the game of Diplomacy, for instance [113], which requires negotiating with other players to form alliances and capture territory. Another possible source of manipulation is reinforcement learning from human feedback (RLHF) [44]. RLHF involves learning a reward function from human feedback to represent a human's preferences, and subsequently training an AI system to optimize that reward function. In general, there may be an incentive for the AI to exert control over the human and their feedback channel so as to maximize reward [93].

In the context of language, RLHF is used to finetune LMs to maximize a human's approval of their behaviour. Without constraints on behaviour, systems trained with RLHF likely have an incentive to obtain human labelers' approval by any means possible, including potentially manipulative avenues. For example, Snoswell and Burgess [150] remark that LMs often seem authoritative even when the information they provide is wrong. A possible explanation is that authoritative language fools human labelers to approve such outputs despite their underlying incorrectness. Chatbots trained with RLHF could also use emojis to appeal to emotions in manipulative ways [166].

Measurement. Work on measuring the incentives and intents of the behaviour of LMs is still preliminary. No existing work applies the CID framework to LMs. While advances in interpretability may help to identify CIDs, this direction remains speculative.

Another line of work has focused on understanding how certain training objectives and environments cause AI systems to generalize differently. Langosco et al. [102] and Shah et al. [146] show that both language models and general RL agents can pursue different goals in out-of-distribution environments even when trained to perfect accuracy on in-distribution environments.

Some recent work has focused on studying the harms and behaviour changes due to LMs. Bender et al. [23], Weidinger et al. [171] outline risks that LMs pose, including informational harms like disinformation. There is a body of work that measures the effects of user interaction with chatbots, in areas like mental health [164], customer service [9], and general assistance [46, 82]. Since there are likely to be domain-dependent manipulation techniques, it would be important to build on this existing work for measuring manipulation.

5 REGULATION OF MANIPULATION

Regulation across different domains apply to human manipulation. In this section, we identify how existing regulations may apply to algorithmic manipulation.

Law. Some manipulation-adjacent acts such as deception or coercion are considered to be sufficiently morally wrong for them to be considered by criminal law. Alternatively, the regulation of certain manipulative practices might have economic justification. In instances where there is an severe asymmetry in power between parties, anti-manipulation regulation can play a role to further social goals such like fairness or the protection of human rights. Anti-manipulation law might therefore appear in contract, tort, competition, market regulatory, consumer or employment law; but as Sunstein [155] notes, such fracturing makes building a common and consistent legal account of manipulation difficult.

Commerce. Calo [33] considers how the trend for extensive data-gathering on individuals makes them more vulnerable to tailored manipulative behaviour. Specifically, digital commerce companies might be able to use finegrained data to limit the consumer's ability to pursue their own interests in a rational manner. He characterises market manipulation as "nudging for profit" and cites the "persuasion profiling" of Kaptein and Duplinsky [92] as one particular example where companies alter their advertising.

---

[2]While this idea is still debated as it involves assuming comparability between different selves [32, 126, 127], we think it nonetheless offers a good starting point.

Willis [175] sees manipulation of consumers as inevitable in the face of AI-enabled systems designed to maximised profit. Unless law and evidential standards are updated, she argues that enforcement will be very difficult. Although intent is not a prerequisite of most state and federal deceptive trading practice law, since it is so difficult to prove, courts still see its proof as a key piece of evidence. This is problematic given the lack of legal precedent concerning intent in algorithms. Further, Willis [175] points to the practical difficulties in proving that a personalised advert is manipulative: typical reasonable person tests are no longer applicable in a world where marketing material for example might be both targeted for *specific* individual at a *specific* point in their day. Organisations that use this type of microtargeting personalization generate so many different user experiences that they might not be able to feasibly monitor them all or recover them when required.

Aside from applications in commerce, microtargeting and related AI-induced manipulation have been discussed as a risk to democratic society [145]. Zuiderveen Borgesius et al. [183] discuss the prospect of tailoring information to boost or decrease voter engagement. Microtargeting is related to hypernudging, which is the use of nudges in a dynamic and pervasive way that is enabled by big data [116, 178]. Nudging [159], which is the design of choice architecture to alter behaviour in a predictable way without changing economic incentives or reducing choice, has long been criticized as potentially being manipulative; for a review of the arguments and counterarguments see Schmidt and Engelen [143]. We note that nudging is actively being pursued in recommender systems [84].

Finance. The spectre of algorithm-led manipulation has already received widespread attention in financial markets. A wide number of financial regulatory laws prohibit a variety of market manipulative practices [135] and algorithmic trading already dominates almost all electronic markets. Unfortunately, a consistent rationale as to why certain trading practices are deemed legal whilst others are not is not forthcoming [47]. Financial regulators following a principles-based approach generally characterise market manipulation as behaviour which gives a false sense of real supply and demand, and by extension price, in a market or benchmark. Market manipulation must be intentional in the US [37], while in the UK intention is not a requirement [14]. As Huang [78] notes, removing intent requirements from regulation, particularly criminal law, is not straightforward.

Regulations designed primarily to regulate human traders may be difficult to enforce in a world where algorithms transact with each other [105]. Bathaee [21] and Scopino [144] both zero in on the intent requirement in proving instances of market manipulation. The view that existing regulations are not sufficient to police market places populated by autonomous learning algorithms is becoming more accepted [16] and solutions are beginning to be mapped out [15] which aim to balance the need to reduce the enforcement gap without unduly chilling AI use in marketplaces.

6 PRACTICAL CHALLENGES FOR FUTURE RESEARCH

The study of manipulation from AI systems presents a number of practical challenges. As shown in Table 2, studies of manipulation can be categorised along each of two axes, making four classes.

The first axis concerns whether the studied system is deployed or simulated. It is extremely difficult for academics and regulators to obtain access to deployed models. While the companies deploying such systems likely have motivated and competent researchers who study manipulation and other problems, institutional barriers can stymie their work, and conflicts of interest may influence crucial decisions. For example, company executives may withhold funding from lines of work deemed too threatening to the company's bottom line [131].

The second axis concerns whether the studied targets of the manipulative system are real or simulated. Simulation has been popular for modelling the effect on users of recommender system [35, 48, 86, 111], but not without criticism [39, 176]. Simulation is cheaper, but has reduced validity, particularly as preference change is not well understood [11, 59, 65]. Efforts exist to address this issue empirically [130] and theoretically [70]. Another advantage of simulation is that the many ethical implications of running manipulation experiments [66] are reduced when the subjects are not human.

Table 1 assesses the difficulties of AI manipulation experimental research. Other than those already mentioned in this section, two further issues exist related to causality. Firstly, as observed in [91], manipulative and adjacent practices are likely to exist simultaneously, so some care needs to be taken to separate them. Secondly, since interaction with any stimuli will change the user, the non-volitional element of that change needs to be measured in order to assess manipulative impact [12]. This challenge has no obvious solution; existing approaches have either attempted to simulate a natural preference evolution [35, 56] or just pretend the user had never interacted with the system [54, 182].

7 CONCLUSION

Although designer intent remains salient, the deployment of opaque and increasingly autonomous systems heightens the importance of a conception of manipulation that can account for manipulation occurring without designer intent. Such manipulation could emerge because it is favoured under the training objective (such as engagement maximization in certain content recommendation settings), or because a model learns to imitate manipulative behavior in its training data (such as manipulative text in language modeling).

We characterized the space of possible definitions of manipulation from AI systems. We analyzed four axes mentioned in prior literature in the context of manipulation by algorithms and other humans: incentives, intent, covertness, and harm. Incentives concern what a system should do to optimize its objective; intent concerns whether a system behaves as if it is reasoning and pursuing an incentive; covertness concerns whether those affected by the system's behaviour meaningfully understand what the system is doing; harm concerns the extent to which the behavior of the AI system negatively affects its users. Although work to operationalize each of these axes exists, fundamental challenges remain.

Manipulation threatens human autonomy [101, 134, 158]. Despite the difficulty of formalizing and measuring manipulation, precautionary action is warranted to anticipate and mitigate potential cases of such behavior. Mitigating actions could include making auditing easier to perform [136, 137], addressing perverse incentives

Challenge	Description	1 LA	2 SPA	3 LUS	4 LSS
Ecological Validity	Simulation of either the user response or the manipulator raises questions about the realism of the simulation. This is particularly acute when simulating humans as the manipulee since this requires modelling their beliefs, behaviour or preference and how it they may change as a result of a manipulative scheme, in addition to exogenous factors.	L	H	H	H
Ethical	Experiments which involve the manipulation of humans are ethically problematic. Experiments which reveal previously unknown vulnerabilities of humans or other systems could constitute info hazards.	M	L	H	L
Access	Owners of systems that might be manipulative have no obvious incentive to allow independent oversight. Data access for researchers is an issue unless they are prepared to build systems to gather and store relevant data themselves.	H	H	-	-
Legality	Conducting research on deployed systems is typically a breach of the standard user agreement → litigation risk.	M	H	-	-
Scale	Large-scale studies are potentially necessary to neutralise effect of confounders.	H	M	H	L
Long Timeframe	Manipulative schemes may only exhibit their effects over long periods of time. This means that efforts to detect it with human users are expensive and trickier to administer. Manipulative effects may be subtle over the typical durations that lab-based user studies take.	H	M	H	L
Measurement	Measuring e.g. preference, belief, or mood change is not straightforward. Behavioural change is easier to measure but will likely not capture all induced change.	H	L	H	L
Stimuli	To measure manipulation in the lab, how should the UX be designed and which stimuli should be used?	-	-	H	M
Baselines	Any interaction with the system will likely change the user. Some of this change is self-induced or desired by the user. It would be wrong to attribute the responsibility for that change to the recommender. This is a puzzle for experiment design – what baseline should be used to measure the presence or absence of manipulative behaviour? E.g. [35, 56] try to estimate 'natural' preference trajectories.	H	-	H	-
Causal Attribution	Manipulative strategies may also be coercive or work in a number of ways simultaneously. How do we attribute behaviour change to manipulation vs other concepts like persuasion or coercion?	H	-	H	-

Table 1: We set out key challenges to a manipulation experiment and rate their difficulty versus the four experiment types described in Table 2. L = Low, M = Medium, H = High, - = N/A

User	(AI) System	
	Real or deployed	Simulated / Toy
Real	1. Live Audit (LA)	3. Lab-based User Study (LUS)
Simulated	2. Sock Puppet Audit (SPA)	4. Lab Simulation Study (LSS)

Table 2: Manipulation study taxonomy.

to build manipulative systems [31], and improving user understanding of AI systems' functioning. Both technical and sociotechnical work to define and measure manipulation should continue, but we should not require certainty before engaging in precautionary and pragmatic mitigations.

ACKNOWLEDGMENTS

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* 24, 4 (Dec. 2013), 956–975. https://doi.org/10.1287/isre.2013.0497 Publisher: INFORMS.

[2] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2018. Effects of Online Recommendations on Consumers' Willingness to Pay. *Information Systems Research* 29, 1 (March 2018), 84–102. https://doi.org/10.1287/isre.2017.0703

[3] M. Mehdi Afsar, Trafford Crump, and Behrouz Far. 2021. Reinforcement learning based recommender systems: A survey. *arXiv:2101.06286 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.06286 arXiv: 2101.06286.

[4] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The Welfare Effects of Social Media. *American Economic Review* 110, 3 (March 2020), 629–676. https://doi.org/10.1257/aer.20190658

[5] Hunt Allcott, Matthew Gentzkow, and Lena Song. 2022. Digital Addiction. *American Economic Review* 112, 7 (July 2022), 2424–2463. https://doi.org/10.1257/aer.20210867

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]* (July 2016). http://arxiv.org/abs/1606.06565 arXiv: 1606.06565.

[7] Jacob Andreas. 2022. Language Models as Agent Models. https://doi.org/10.48550/arXiv.2212.01681 arXiv:2212.01681 [cs].

[8] Stuart Armstrong. 2015. Motivated Value Selection for Artificial Agents. (2015).

[9] Muhammad Ashfaq, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. 2020. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics* 54 (Nov. 2020), 101473. https://doi.org/10.1016/j.tele.2020.101473

[10] Hal Ashton. 2022. Definitions of Intent Suitable for Algorithms. *Artificial Intelligence and Law* (July 2022). https://doi.org/10.1007/s10506-022-09322-x

[11] Hal Ashton and Matija Franklin. 2022. The Problem of Behaviour and Preference Manipulation in AI Systems. In *The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)*.

[12] Hal Ashton and Matija Franklin. 2022. Solutions to Preference Manipulation in Recommender Systems Require Knowledge of Meta-Preferences. http://arxiv.org/abs/2209.11801 arXiv:2209.11801 [cs].

[13] Association for Computing Machinery (ACM). 2019. "Reinforcement Learning for Recommender Systems: A Case Study on Youtube," by Minmin Chen. https://www.youtube.com/watch?v=HEqQ2_1XRTs

[14] Financial Conduct AuthorityA. 2016. FCA Handbook: MAR 1 Market Abuse. https://www.handbook.fca.org.uk/handbook/MAR.pdf

[15] Alessio Azzutti. 2022. AI-driven Market Manipulation and Limits of the EU Law Enforcement Regime to Credible Deterrence. *Computer Law & Security review* 45 (Jan. 2022). https://doi.org/10.2139/ssrn.4026468

[16] Alessio Azzutti, Wolf-Georg Ringe, and H. Siegfried Stiehl. 2021. Machine Learning, Market Manipulation and Collusion on Capital Markets: Why the. *University of Pennsylvania journal of international law* 43, 1 (2021). https://doi.org/10.2139/ssrn.3788872

[17] Hui Bai. 2023. Artificial Intelligence Can Persuade Humans. (2023).

[18] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445717

[19] Marcia Baron. 2014. The Mens Rea and Moral Status of Manipulation. In *Manipulation: Theory and Practice*, Christian Coons and Michael Weber (Eds.). Oxford University Press, 0. https://doi.org/10.1093/acprof:oso/9780199338207.003.0005

[20] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C. Mitchell. 2020.

Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376518

[21] Yavar Bathaee. 2018. The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law and Technology* 31, 2 (2018), 890–938.

[22] Omer Ben-Porat and Moshe Tennenholtz. 2018. A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/a9a1d5317a33ae8cef33961c34144f84-Paper.pdf

[23] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[24] Jonah Berger and Katherine L. Milkman. 2012. What Makes Online Content Viral? *Journal of Marketing Research* 49, 2 (April 2012), 192–205. https://doi.org/10.1509/jmr.10.0353

[25] J. S. Blumenthal-Barby and Hadley Burroughs. 2012. Seeking Better Health Care Outcomes: the Ethics of Using the "Nudge". *The American journal of bioethics: AJOB* 12, 2 (2012), 1–10. https://doi.org/10.1080/15265161.2011.634481

[26] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[27] Harriet Braiker. 2003. *Who's Pulling Your Strings?: How to Break the Cycle of Manipulation and Regain Control of Your Life: How to Break the Cycle of Manipulation and Regain Control of Your Life*. McGraw Hill Professional. Google-Books-ID: dGwgiQvyeq0C.

[28] Michael Bratman. 1987. Intention, plans, and practical reason. https://philpapers.org/rec/BRAIPA

[29] Harry Brignull. 2018. Deceptive Design - User Interfaces Crafted to Trick You. https://www.deceptive.design/

[30] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=ETKGuby0hcs

[31] Qingpeng Cai, Shuchang Liu, Xueliang Wang, Tianyou Zuo, Wentao Xie, Bin Yang, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Reinforcing User Retention in a Billion Scale Short Video Recommender System. http://arxiv.org/abs/2302.01724 arXiv:2302.01724 [cs].

[32] Agnes Callard. 2018. *Aspiration: The Agency of Becoming*. Oxford University Press, Oxford, New York.

[33] M. Ryan Calo. 2014. Digital Market Manipulation. *George Washington Law Review* 82, 4 (2014), 996–1051. https://doi.org/10.2139/ssrn.2309703

[34] Ryan Carey, Eric Langlois, Tom Everitt, and Shane Legg. 2020. The Incentives that Shape Behaviour. *arXiv:2001.07118 [cs]* (Jan. 2020). http://arxiv.org/abs/2001.07118 arXiv: 2001.07118.

[35] Micah Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. Estimating and Penalizing Induced Preference Shifts in Recommender Systems. *Proceedings of machine learning research* 162 (2022), 2686–2708.

[36] Thomas L. Carson. 2010. *Lying and Deception: Theory and Practice*. Oxford University Press, Oxford ; New York. OCLC: ocn464581525.

[37] CFTC. 2013. *Antidisruptive Practices Authority Interpretative Guidance and Policy Statement*. Technical Report RIN 3038-AD96. Commodity Futures Trading Commission. https://www.federalregister.gov/documents/2013/05/28/2013-12365/antidisruptive-practices-authority

[38] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. https://doi.org/10.48550/arXiv.2302.10329 arXiv:2302.10329 [cs].

[39] Allison J. B. Chaney. 2021. Recommendation System Simulations: A Discussion of Two Key Challenges. https://doi.org/10.48550/arXiv.2109.02475

[40] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *Proceedings of the 12th ACM Conference on Recommender Systems* (Sept. 2018), 224–232. https://doi.org/10.1145/3240323.3240370 arXiv:1710.11214.

[41] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed Chi. 2020. Top-K Off-Policy Correction for a REINFORCE Recommender System. *arXiv:1812.02353 [cs, stat]* (Nov. 2020). http://arxiv.org/abs/1812.02353 arXiv: 1812.02353.

[42] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and others. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).

[43] Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. *Eliciting Latent Knowledge*. Technical Report. Alignment Research Center. https://ai-alignment.com/eliciting-latent-knowledge-f977478608fc

[44] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *arXiv:1706.03741 [cs, stat]* (July 2017). http://arxiv.org/abs/1706.03741 arXiv: 1706.03741.

[45] Thomas Christiano. 2022. Algorithms, Manipulation, and Democracy. *Canadian Journal of Philosophy* 52, 1 (Jan. 2022), 109–124. https://doi.org/10.1017/can.2021.29 Publisher: Cambridge University Press.

[46] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the Shades of the Uncanny Valley: An Experimental Study of Human–Chatbot Interaction. *Future Generation Computer Systems* 92 (March 2019), 539–548. https://doi.org/10.1016/j.future.2018.01.055

[47] Ricky Cooper, Michael Davis, and Ben Van Vliet. 2016. The Mysterious Ethics of High-Frequency Trading. *Business Ethics Quarterly* 26, 1 (Jan. 2016), 1–22. https://doi.org/10.1017/beq.2015.41

[48] Mihaela Curmei, Andreas A. Haupt, Benjamin Recht, and Dylan Hadfield-Menell. 2022. Towards Psychologically-Grounded Dynamic Preference Models. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 35–48.

[49] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. https://doi.org/10.48550/arXiv.2012.08630 arXiv:2012.08630 [cs].

[50] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. 2019. Where's the Reward? *International Journal of Artificial Intelligence in Education* 29, 4 (Dec. 2019), 568–620. https://doi.org/10.1007/s40593-019-00187-x

[51] Robert Epstein and Ronald E. Robertson. 2015. The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections. *Proceedings of the National Academy of Sciences* 112, 33 (Aug. 2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112 Publisher: Proceedings of the National Academy of Sciences.

[52] Charles Evans and Atoosa Kasirzadeh. 2021. User Tampering in Reinforcement Learning Recommender Systems. *arXiv:2109.04083 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.04083 arXiv: 2109.04083.

[53] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and Governing AI that does not Lie. *arXiv:2110.06674 [cs]* (Oct. 2021). http://arxiv.org/abs/2110.06674 arXiv: 2110.06674.

[54] Tom Everitt, Ryan Carey, Eric Langlois, Pedro A. Ortega, and Shane Legg. 2021. Agent Incentives: A Causal Perspective. arXiv:2102.01685.

[55] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *arXiv:1908.04734 [cs]* (March 2021). http://arxiv.org/abs/1908.04734 arXiv: 1908.04734.

[56] Sebastian Farquhar, Ryan Carey, and Tom Everitt. 2022. Path-Specific Objectives for Safer Agent Incentives. *arXiv:2204.10018 [cs, stat]* (April 2022). http://arxiv.org/abs/2204.10018 arXiv: 2204.10018.

[57] Brian J Fogg. 1998. Captology: the Study of Computers as Persuasive Technologies. In *CHI 98 Conference Summary on Human Factors in Computing Systems*. 385.

[58] Brian J Fogg. 2003. *Persuasive Technology*. Elsevier. https://doi.org/10.1016/B978-1-55860-643-2.X5000-8

[59] Matija Franklin, Hal Ashton, Rebecca Gorman, and Stuart Armstrong. 2022. Recognising the Importance of Preference Change: A Call for a Coordinated Multidisciplinary Research Effort in the Age of AI. *arXiv:2203.10525 [cs]* (March 2022). http://arxiv.org/abs/2203.10525 arXiv: 2203.10525.

[60] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533229

[61] Jason Gauci, Edoardo Conti, Yitao Liang, Kittipat Virochsiri, Yuchen He, Zachary Kaden, Vivek Narayanan, Xiaohui Ye, Zhengxing Chen, and Scott Fujimoto. 2019. Horizon: Facebook's Open Source Applied Reinforcement Learning Platform. *arXiv:1811.00260 [cs, stat]* (Sept. 2019). http://arxiv.org/abs/1811.00260 arXiv: 1811.00260.

[62] Charles Goodhart. 1975. Problems of Monetary Management: the UK Experience in Papers in Monetary Economics. *Monetary Economics* 1 (1975).

[63] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3174108

[64] Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. 2023. Susceptibility to Influence of Large Language Models. http://arxiv.org/abs/2303.06074 arXiv:2303.06074 [cs].

[65] Till Grüne-Yanoff and Sven Ove Hansson (Eds.). 2009. *Preference change: approaches from philosophy, economics and psychology*. Number v. 42 in Theory and decision library. Series A, Philosophy and methodology of the social sciences. Springer, Dordrecht ; London. OCLC: ocn321018474.

[66] Blake Hallinan, Jed R Brubaker, and Casey Fiesler. 2020. Unexpected Expectations: Public Reaction to the Facebook Emotional Contagion Study. *New Media & Society* 22, 6 (June 2020), 1076–1094. https://doi.org/10.1177/1461444819876944

[67] Joseph Y. Halpern and Max Kleiman-Weiner. 2018. Towards Formal Definitions of Blameworthiness, Intention, and Moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, New Orleans, Louisiana, USA, 1853–1860.

[68] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting Consumption towards Diverse Content on Music Streaming Platforms. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, Virtual Event Israel, 238–246. https://doi.org/10.1145/3437963.3441775

[69] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS '16)*. Association for Computing Machinery, New York, NY, USA, 111–122.

[70] Adrian Haret and Johannes Peter Wallner. 2022. An Axiomatic Approach to Revising Preferences. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 5 (June 2022), 5676–5683. https://doi.org/10.1609/aaai.v36i5.20509

[71] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

[72] D. Heckerman and R. Shachter. 1995. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research* 3 (Dec. 1995), 405–430. https://doi.org/10.1613/jair.202

[73] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. 2018. Causal Structure Learning. *Annual Review of Statistics and Its Application* 5, 1 (2018), 371–391. https://doi.org/10.1146/annurev-statistics-031017-100630 _eprint: https://doi.org/10.1146/annurev-statistics-031017-100630.

[74] Jacob B. Hirsh, Sonia K. Kang, and Galen V. Bodenhausen. 2012. Personalized Persuasion: Tailoring Persuasive Appeals to Recipients' Personality Traits. *Psychological Science* 23, 6 (June 2012), 578–581. https://doi.org/10.1177/0956797611436349 Publisher: SAGE Publications Inc.

[75] Joey Hong, Anca Dragan, and Sergey Levine. 2023. Learning to Influence Human Behavior with Offline Reinforcement Learning. https://doi.org/10.48550/arXiv.2303.02265 arXiv:2303.02265 [cs].

[76] Yubo Hou, Dan Xiong, Tonglin Jiang, Lily Song, and Qi Wang. 2019. Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 13, 1 (Feb. 2019). https://doi.org/10.5817/CP2019-1-4

[77] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 259–268.

[78] (Robin) Hui Huang. 2009. Redefining Market Manipulation in Australia: The Role of an Implied Intent Element. *Companies and Securities Law Journal* 27 (April 2009). https://papers.ssrn.com/abstract=1376209

[79] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2021. Algorithmic Amplification of Politics on Twitter. *arXiv:2110.11010 [cs]* (Oct. 2021). http://arxiv.org/abs/2110.11010 arXiv: 2110.11010.

[80] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. 2019. Human-level Performance in 3D Multiplayer Games with Population-Based Reinforcement Learning. *Science* 364, 6443 (May 2019), 859–865. https://doi.org/10.1126/science.aau6249 Publisher: American Association for the Advancement of Science.

[81] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. 2021. Alternative Microfoundations for Strategic Classification. In *ICML*. http://arxiv.org/abs/2106.12705 arXiv: 2106.12705.

[82] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. https://doi.org/10.1145/3544548.3581196 arXiv:2302.00560 [cs].

[83] Janus. 2022. Simulators. https://generative.ink/posts/simulators/

[84] Mathias Jesse and Dietmar Jannach. 2021. Digital Nudging with Recommender Systems: Survey and Future Directions. *Computers in Human Behavior Reports* 3 (Jan. 2021), 100052. https://doi.org/10.1016/j.chbr.2020.100052

[85] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* (Nov. 2022). https://doi.org/10.1145/3571730 Just Accepted.

[86] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Jan. 2019), 383–390. https://doi.org/10.1145/3306618.3314288 arXiv: 1902.10730.

[87] Eric J. Johnson and Daniel Goldstein. 2003. Do Defaults Save Lives? *Science* 302, 5649 (Nov. 2003), 1338–1339. https://doi.org/10.1126/science.1091721 Publisher: American Association for the Advancement of Science.

[88] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. https://doi.org/10.48550/arXiv.2207.05221 arXiv:2207.05221 [cs].

[89] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. https://doi.org/10.48550/arXiv.2206.15475 arXiv:2206.15475 [cs, stat].

[90] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian Persuasion. *American Economic Review* 101, 6 (Oct. 2011), 2590–2615. https://doi.org/10.1257/aer.101.6.2590

[91] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. 2018. Coercion and Deception in Persuasive Technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*. CEUR-WS, 38–49.

[92] Maurits Kaptein and Steven Duplinsky. 2013. Combining Multiple Influence Strategies to Increase Consumer Compliance. *International Journal of Internet Marketing and Advertising* 8, 1 (2013), 32. https://doi.org/10.1504/IJIMA.2013.056586

[93] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. *arXiv:2103.14659 [cs]* (March 2021). http://arxiv.org/abs/2103.14659 arXiv: 2103.14659.

[94] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2022. Discovering Agents. https://doi.org/10.48550/arXiv.2208.08345 arXiv:2208.08345 [cs].

[95] Poruz Khambatta, Shwetha Mariadassou, Joshua Morris, and S Christian Wheeler. 2022. Targeting Recommendation Algorithms to Ideal Preferences Makes Users Better Off. (2022).

[96] Jon Kleinberg and Manish Raghavan. 2019. How Do Classifiers Induce Agents to Invest Effort Strategically?. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC '19)*. Association for Computing Machinery, New York, NY, USA, 825–844. https://doi.org/10.1145/3328526.3329584

[97] Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. http://arxiv.org/abs/2302.02083 arXiv:2302.02083 [cs].

[98] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. 2019. Penalizing Side Effects Using Stepwise Relative Reachability.

arXiv:1806.01186 [cs, stat] (March 2019). http://arxiv.org/abs/1806.01186 arXiv: 1806.01186.

[99] Ilan Kremer, Yishay Mansour, and Motty Perry. 2014. Implementing the "Wisdom of the Crowd". *Journal of Political Economy* 122, 5 (2014), 988 – 1012. https://econpapers.repec.org/article/ucpjpolec/doi_3a10.1086_2f676597.htm Publisher: University of Chicago Press.

[100] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden Incentives for Auto-Induced Distributional Shift.

[101] Arto Laitinen and Otto Sahlgren. 2021. AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence* 4 (2021). https://www.frontiersin.org/articles/10.3389/frai.2021.705164

[102] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12004–12019. https://proceedings.mlr.press/v162/langosco22a.html

[103] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *The Eleventh International Conference on Learning Representations.* https://openreview.net/forum?id=DeG07_TcZvT

[104] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.07958 arXiv: 2109.07958.

[105] Tom C. W. Lin. 2017. The New Market Manipulation. *Emory Law Journal* 66 (July 2017). https://papers.ssrn.com/abstract=2996896

[106] David Lindner, Kyle Matoba, and Alexander Meulemans. 2021. Challenges for Using Impact Regularizers to Avoid Negative Side Effects. *arXiv:2101.12509 [cs]* (Feb. 2021). http://arxiv.org/abs/2101.12509 arXiv: 2101.12509.

[107] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* 13, 1 (March 2021), 43–109. https://doi.org/10.1093/jla/laaa006

[108] James Edwin Mahon. 2016. The Definition of Lying and Deception. In *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/lying-definition/

[109] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. http://arxiv.org/abs/2301.06627 arXiv:2301.06627 [cs].

[110] David Manheim and Scott Garrabrant. 2019. Categorizing Variants of Goodhart's Law. *arXiv:1803.04585 [cs, q-fin, stat]* (Feb. 2019). http://arxiv.org/abs/1803.04585 arXiv: 1803.04585.

[111] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. *arXiv:2007.13019 [cs]* (July 2020). http://arxiv.org/abs/2007.13019 arXiv: 2007.13019.

[112] Jörg Meibauer. 2005. Lying and Falsely Implicating. *Journal of Pragmatics* 37, 9 (Sept. 2005), 1373–1399. https://doi.org/10.1016/j.pragma.2004.12.007

[113] Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. *Science* 378, 6624 (Dec. 2022), 1067–1074. https://doi.org/10.1126/science.ade9097 Publisher: American Association for the Advancement of Science.

[114] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. 2023. Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media. https://doi.org/10.48550/arXiv.2305.16941 arXiv:2305.16941 [cs].

[115] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 230–239.

[116] Stuart Mills. 2022. Finding the 'Nudge' in Hypernudge. *Technology in Society* 71 (Nov. 2022), 102117. https://doi.org/10.1016/j.techsoc.2022.102117

[117] Kevin Munger and Joseph Phillips. 2020. Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics* (Oct. 2020), 1940161220964767. https://doi.org/10.1177/1940161220964767 Publisher: SAGE Publications Inc.

[118] Maciej Musiał. 2022. Can We Design Artificial Persons without Being Manipulative? *AI & SOCIETY* (Oct. 2022). https://doi.org/10.1007/s00146-022-01575-z

[119] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey Research in HCI. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). Springer, New York, NY, 229–266. https://doi.org/10.1007/978-1-4939-0378-8_10

[120] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and others. 2021. Webgpt: Browser-Assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332* (2021).

[121] Robert Noggle. 2022. The Ethics of Manipulation. In *The Stanford Encyclopedia of Philosophy* (summer 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/

[122] APA Dictionary of Psychology. 2023. Definition of manipulation. https://dictionary.apa.org/manipulation

[123] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread* (2022).

[124] Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. http://arxiv.org/abs/2304.03279 arXiv:2304.03279 [cs].

[125] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. http://arxiv.org/abs/2308.14752 arXiv:2308.14752 [cs].

[126] L. A. Paul. 2014. *Transformative experience* (1st ed ed.). Oxford University Press, Oxford. OCLC: ocn872342141.

[127] L. A. Paul. 2022. Choosing for Changing Selves. *The Philosophical Review* 131, 2 (April 2022), 230–235. https://doi.org/10.1215/00318108-9554756

[128] Amalie Brogaard Pauli, Leon Derczynski, and Ira Assent. 2022. Modelling Persuasion through Misuse of Rhetorical Appeals. (2022).

[129] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR.

[130] Fabíola S. F. Pereira, João Gama, Sandra de Amo, and Gina M. B. Oliveira. 2018. On Analyzing User Preference Dynamics with Temporal Social Networks. *Machine Learning* 107, 11 (Nov. 2018), 1745–1773. https://doi.org/10.1007/s10994-018-5740-2

[131] Billy Perrigo. 2021. How Frances Haugen's Team Forced a Facebook Reckoning. *Time* (Oct. 2021). https://time.com/6104899/facebook-reckoning-frances-haugen/

[132] Richard Pettigrew. 2019. *Choosing for Changing Selves* (1 ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198814962.001.0001

[133] Richard Pettigrew. 2022. Nudging for Changing Selves. *SSRN Electronic Journal* (2022). https://doi.org/10.2139/ssrn.4025214

[134] Carina Prunkl. 2022. Human Autonomy in the Age of Artificial Intelligence. *Nature Machine Intelligence* 4, 2 (Feb. 2022), 99–101. https://doi.org/10.1038/s42256-022-00449-9 Number: 2 Publisher: Nature Publishing Group.

[135] Tālis Putniņš. 2020. An Overview of Market Manipulation. In *Corruption and Fraud in Financial Markets* (1st ed.), Carol Alexander and Douglas Cumming (Eds.). John Wiley & Sons Inc., United States, 13–44.

[136] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618.3314244

[137] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534181

[138] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

[139] Manoel Horta Ribeiro, Veniamin Veselovsky, and Robert West. 2023. The Amplification Paradox in Recommender Systems. http://arxiv.org/abs/2302.11225 arXiv:2302.11225 [cs].

[140] Jonathan Richens, Rory Beard, and Daniel H. Thompson. 2022. Counterfactual Harm. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=zkQho-Jxky9

[141] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. (2014), 23.

[142] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint*

arXiv:2302.04761 (2023).

[143] Andreas T. Schmidt and Bart Engelen. 2020. The Ethics of Nudging: An Overview. *Philosophy Compass* 15, 4 (April 2020). https://doi.org/10.1111/phc3.12658

[144] Gregory Scopino. 2015. Do Automated Trading Systems Dream of Manipulating the Price of Futures contracts? Policing Markets for Improper Trading Practices by Algorithmic Robots. *Florida Law Review* 67 (2015), 221.

[145] Caroline Serbanescu. 2021. Why Does Artificial Intelligence Challenge Democracy? A Critical Analysis of the Nature of the Challenges Posed by AI-Enabled Manipulation. *Copenhagen journal of legal studies* 5, 1 (2021), 105–128. https://ssrn.com/abstract=4033258

[146] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. https://doi.org/10.48550/arXiv.2210.01790 arXiv:2210.01790 [cs].

[147] caroline sinders. 2022. What's In a Name? https://medium.com/@carolinesinders/whats-in-a-name-unpacking-dark-patterns-versus-deceptive-design-e96068627ec4

[148] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Hacking. http://arxiv.org/abs/2209.13085 arXiv:2209.13085 [cs, stat].

[149] David Horton Smith. 1967. Correcting for Social Desirability Response Sets in Opinion-Attitude Survey Research. *The Public Opinion Quarterly* 31, 1 (1967), 87–94. https://www.jstor.org/stable/2746886 Publisher: [Oxford University Press, American Association for Public Opinion Research].

[150] Aaron J. Snoswell and Jean Burgess. 2022. The Galactica AI Model was Trained on Scientific Knowledge – but it Spat Out Alarmingly Plausible Nonsense. http://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nonsense-195445

[151] Barry M. Staw. 1976. Knee-Deep in the Big Muddy: a Study of Escalating Commitment to a Chosen Course of Action. *Organizational Behavior and Human Performance* 16, 1 (June 1976), 27–44. https://doi.org/10.1016/0030-5073(76)90005-2

[152] Jacob Steinhardt. 2023. Emergent Deception and Emergent Optimization. https://bounded-regret.ghost.io/emergent-deception-optimization/

[153] Jonathan Stray, Steven Adler, and Dylan Hadfield-Menell. 2021. What are you optimizing for? Aligning Recommender Systems with Human Values. (2021), 7.

[154] Michael Strevens. 2020. *The Knowledge Machine: How Irrationality Created Modern Science.* Liveright Publishing.

[155] Cass R. Sunstein. 2021. Manipulation As Theft. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3880048

[156] Daniel Susser. 2019. Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19).* Association for Computing Machinery, New York, NY, USA, 403–408. https://doi.org/10.1145/3306618.3314286

[157] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Online Manipulation: Hidden Influences in a Digital World. *Geo. L. Tech. Rev.* 4 (2019), 1. Publisher: HeinOnline.

[158] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, Autonomy, and Manipulation. *Internet Policy Review* 8, 2 (June 2019). https://papers.ssrn.com/abstract=3420747

[159] Richard H. Thaler and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth and Happiness* (revised edition, new international edition ed.). Penguin Books, London New York Toronto Dublin Camberwell New Delhi Rosedale Johannesburg.

[160] Luke Thorburn. 2022. How Platform Recommenders Work. https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a

[161] Luke Thorburn, Jonathan Stray, and Priyanjana Bengani. 2022. What Will "Amplification" Mean in Court? https://techpolicy.press/what-will-amplification-mean-in-court/?curius=1684

[162] Twitter. 2023. Twitter's Recommendation Algorithm. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

[163] Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. http://arxiv.org/abs/2302.08399 arXiv:2302.08399 [cs].

[164] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. 2019. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie* 64, 7 (July 2019), 456–464. https://doi.org/10.1177/0706743719828977

[165] James Vincent. 2023. Microsoft's Bing is an emotionally manipulative liar, and people love it. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

[166] Carissa Véliz. 2023. Chatbots Shouldn't Use Emojis. *Nature* 615, 7952 (March 2023), 375–375. https://doi.org/10.1038/d41586-023-00758-y Bandiera_abtest: a Cg_type: World View Number: 7952 Publisher: Nature Publishing Group Subject_term: Ethics, Society, Machine learning, Technology.

[167] Francis Rhys Ward. 2022. On Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios. (2022).

[168] Francis Rhys Ward, Tom Everitt, Francesca Toni, and Francesco Belardinelli. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. (2023).

[169] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2022. A Causal Perspective on AI Deception in Games. (2022).

[170] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD

[171] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM. https://doi.org/10.1145/3531146.3533088

[172] Benjamin Weissman and Marina Terkourafi. 2019. Are False Implicatures Lies? An Empirical Investigation. *Mind & Language* 34, 2 (2019), 221–246. https://doi.org/10.1111/mila.12212 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12212.

[173] Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. 2021. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *Wall Street Journal* (Sept. 2021). https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

[174] Nicole Wetsman. 2021. Facebook's Whistleblower Report Confirms what Researchers Have Known for Years. *The Verge* (Oct. 2021). https://www.theverge.com/2021/10/6/22712927/facebook-instagram-teen-mental-health-research

[175] Lauren E. Willis. 2020. Deception by Design. *Harvard journal of law and technology* 34, 1 (Aug. 2020). https://papers.ssrn.com/abstract=3694575

[176] Amy A. Winecoff, Matthew Sun, Eli Lucherini, and Arvind Narayanan. 2021. Simulation as Experiment: An Empirical Critique of Simulation Research on Recommender Systems. http://arxiv.org/abs/2107.14333 arXiv:2107.14333 [cs].

[177] Allen W. Wood. 2014. Coercion, Manipulation, Exploitation. In *Manipulation: theory and practice.* Oxford University Press, Oxford ; New York. DOI:10.1093/acprof:oso/9780199338207.003.0002

[178] Karen Yeung. 2017. 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society* 20, 1 (Jan. 2017), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

[179] Savvas Zannettou, Sotirios Chatzis, Kostantinos Papadamou, and Michael Sirivianos. 2018. The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube. In *2018 IEEE Security and Privacy Workshops (SPW).* 63–69. https://doi.org/10.1109/SPW.2018.00018

[180] Tal Z. Zarsky. 2019. Privacy and Manipulation in the Digital Age. *Theoretical Inquiries in Law* 20, 1 (March 2019), 157–188. https://doi.org/10.1515/til-2019-0006

[181] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[182] Zhengbang Zhu, Rongjun Qin, Junjie Huang, Xinyi Dai, Yang Yu, Yong Yu, and Weinan Zhang. 2022. Understanding or Manipulation: Rethinking Online Performance Gains of Modern Recommender Systems. http://arxiv.org/abs/2210.05662 arXiv:2210.05662 [cs].

[183] Frederik Zuiderveen Borgesius, Judith Moeller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balázs Bodó, and Claes H. de Vreese. 2018. Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review* 14, 1 (Feb. 2018), 82–96. https://papers.ssrn.com/abstract=3128787