Characterizing Manipulation from AI Systems

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

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

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

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

ABSTRACT

Manipulation is a common concern in many domains, such as social media, advertising, and chatbots. As AI systems mediate more of our interactions with the world, 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 manipulation in the context of AI systems. Firstly, we build upon prior literature on manipulation from other fields and characterize the space of possible notions of manipulation, which we find to depend upon the concepts of incentives, intent, harm, and covertness. We review proposals on how to operationalize each factor. Second, we propose a definition of manipulation based on our characterization: a system is manipulative *if it acts as if it were pursuing an incentive to change a human (or another agent) intentionally and covertly*. Third, we discuss the connections between manipulation and related concepts, such as deception and coercion. Finally, we contextualize our operationalization of manipulation in some applications. Our overall assessment is that 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. We argue that such manipulation poses a significant threat to human autonomy, suggesting that precautionary actions to mitigate it are warranted.

KEYWORDS

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

ACM Reference Format:
Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing Manipulation from AI Systems. In *Proceedings of Under review*. ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

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

Under review, XX, XX

ACM ISBN 978-1-4503-XXXX-X/18/06…$0
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A characteristic of intelligent agents is the ability to change the environment around them to further their own 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 [76, 158], 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 [39, 56, 93, 101, 166].

While designers and operators might engage in manipulative behaviour with the aid of AI systems [48, 118], we concentrate on the ways in which an AI system may itself be manipulative. This distinction is not to say that designer or operator manipulation is unimportant (e.g. disinformation campaigns). Rather, our focus is motivated by the increasingly evident fact that systems exhibit capabilities that designers do not necessarily foresee or intend [30, 42, 169]. Our notion of manipulation from AI systems can be considered a subset of algorithmic manipulation, which we take to encompass both concerns.

There are two main reasons why we can reasonably expect AI systems to manipulate humans absent designer intent. First, AI systems learn to imitate human manipulation when their training data contain examples of manipulative behaviour. Language models trained on internet data seem capable of reproducing the persuasive and manipulative tactics of humans [20, 65, 164]. Second, manipulative behaviour may be unintentionally optimal for objective functions which we provide to machine-learning (ML) systems. As a real-world example, consider a recommender system trained to maximize user watch time over a session. Maximizing watch time might involve influencing the user to start watching a 10-part video series that they will feel compelled to finish because of cognitive biases such as sunk cost fallacy [152], rather than because of actual value derived.

As it stands, there is limited literature regarding manipulation from AI systems. We believe there are two reasons for this state of affairs. Firstly, in common with many other behaviours or mental states with folk-definitions, it is difficult to construct a definition of manipulation which is both general enough to cover a wide variety of cases and specific enough to be practically implementable [48, 121]. Secondly, the testing of any putative definition is beset

with difficulties. Monitoring the impact of deployed systems *in situ* is not easy without the express permission of the system (and data) owners. Since the conclusions of such research might be reputationally negative, this permission is rarely forthcoming.[1] 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 manipulativeness.

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 synthesize our characterization to propose a definition for manipulation. We discuss how our definition applies to a variety of systems and discuss problems with its usage. Third, we compare our definition to notions adjacent to manipulation. Third, we discuss the operationalization of manipulation in the context of recommender systems and language models. 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 on prior literature concerning manipulation, we characterize the space of possible notions of manipulation from AI systems. Our characterization depends upon four axes: incentives, intent, covertness, and harm.

2.1 Incentives

Our first axis is whether the system has **incentives** for influence: that is, incentives to change a human's behaviour (which in turn, will likely involve changing their beliefs, preferences, or psychological state more broadly). 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 may have incentives to influence user behaviour so as to make them more predictable [39, 101].

Incentives in Prior Definitions of Manipulation. Some definitions of manipulation involve a benefit to the manipulator [31, 123]. 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 requires that the response of the human benefits the AI system in some way.

Operationalizing Incentives. The standard toolkit to analyze incentives of AI systems are causal influence diagrams (CIDs) [54, 56, 74]. Using the notation from Everitt et al. [56], 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. CIDs allow us to talk about **instrumental control incentives** – a formalization of the idea that an AI system chooses a certain behaviour because that behaviour is instrumental to achieving the system's goal. Evans and Kasirzadeh [54] apply their framework to a simple content recommendation example to show how RL systems have incentives to influence user preferences. We provide an example of a CID in Figure 1.

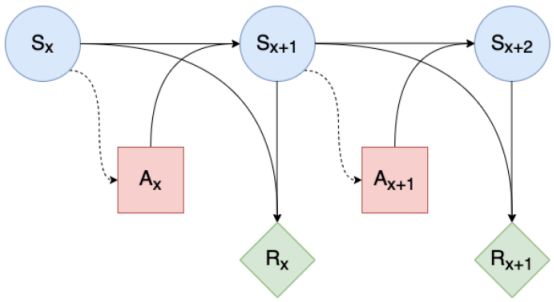

Figure 1: An example of a causal influence diagram (CID) [56], from Figure 1 of Evans and Kasirzadeh [54]. The CID models a content recommendation system that decides which posts A_x at time x to show to the user, based on the user's state S_x. The system receives reward R_x after its action. If the system optimizes for the sum of rewards, it would have an incentive at time x to influence future states S_{x+k} to make it easier to obtain rewards.

An AI system's implicit or explicit causal model may be inaccurate with respect to the underlying causal model of the world, whether by design or suboptimality. For example, a designer may endow an AI system with a causal model in which an AI system's predictions have no causal influence on certain parts of a user's state [38, 54]. Alternatively, a model may learn to reason incorrectly because of spurious correlations during training [103].

Given an AI system's CID, one way to operationalize **incentive** is to use the notion of instrumental control incentive in Everitt et al. [56]. 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. Intuitively, 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 [54, 56, 101]).

One may want to **remove incentives** to influence humans if we are concerned that such influence would be manipulative or otherwise harmful. Incentives can be removed by eliminating the ability of the system to have such influence, or (potentially) by changing the utility function. However, removing incentives to influence users may be intractable, in which case designers might instead **hide incentives** by designing the system to ignore them, for instance, by specifying an (inaccurate) causal model in which an

[1] Perhaps companies do privately monitor the impact of their systems on their users. However, such knowledge could also become incriminating, leading to incentives not to conduct investigations in the first place [173, 174].

AI system's predictions have no causal influence on certain parts of a user's state [38, 58, 101].

Challenges. A disadvantage of the CID framework is that it requires constructing and reasoning about causal graphs. It may often be ambiguous or counterintuitive to determine the correct CID nodes and causal relationships which correspond to a specific training setup, as we explore in Section 3.2. Interpretability tools may help to provide the primitives upon which nodes in a CID can be constructed. For example, Jaderberg et al. [81] 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. Moreover, removing or hiding incentives often comes at the expense of reducing the capability of the system in ways that might render it less useful, as discussed in Section 2.4.

Other Considerations. Any analysis of which incentives exist implicitly depends upon the power of a system to influence humans. Systems with restricted action spaces or 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 change humans will often have incentives to change them, as AI systems' rewards generally depend on humans' actions.

Optimization over long-horizons seems to provide more opportunities for manipulation. For example, a recommender system likely cannot radicalize a user in only one timestep, but a sufficiently capable system optimizing over many timesteps could plan to shift a user over the course of many actions. If such a plan is advantageous for the training objective (e.g., now the user will engage more predictably with such content, leading to higher reward), we should not be surprised to discover if the system executes the plan in practice, subject to difficulties the system may have in finding such strategies.

2.2 Intent

Even when a system has an incentive to influence humans, such an incentive might not be pursued due to limited data, insufficient training, low capacity, or simply due to chance. The notion of *intent* to influence can help to distinguish systems that can be expected to pursue incentives reliably. We emphasize that by referring to an AI system's intent, we are making no statement about algorithmic theory of mind or moral status. We are emphatically *not* absolving designers of the responsibility of designing safe 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. This definition heavily intersects with other definitions of intent for AI systems [13, 68]. We want to distinguish between cases in which the system behaves in a manipulative way incidentally (e.g. by random chance) from when the system's behavior being part of a systematic pattern of manipulation for the purpose of causing a downstream outcome.

Our notion of intent (which is independent of the intent of the designer) emphasizes that it is possible to unintentionally create systems that engage in goal-directed manipulative behavior. A subtler advantage of such definition of intent is that it allows for grounding the notion in a fully behavioral lens, which is agnostic to the actual computational process of the system: e.g. a lookup table could also be understood as engaging in reasoning if it contained a optimal planner's outputs, as discussed in Table 1.

Intent in Prior Definitions of Manipulation. Some definitions of manipulation involve an intent on the part of the manipulator to engage in manipulation [121]. Susser et al. [157] 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 [22] 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] avoids 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 here detail a couple of promising approaches.

Ashton [13] provides several definitions of intent for AI systems that are inspired by criminal law. According to their basic definition of intent, an AI system intends a result through an action if (i) alternative actions exist, (ii) the AI system is capable of observing when the result occurs, (iii) the AI system foresees that the action causes the result, and (iv) the result is beneficial for the AI system. The first and second criteria seem easy to establish. The third and fourth criteria would likely be easier to satisfy with access to the agent's utility function and world model. However, it might be possible to assess those criteria without the utility function or world model, such as through techniques accessing model internals [34, 124] or even asking language models themselves [89]. We expand more upon this possible direction in Section 6.

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. [93] 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 since 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 aware of the *specific ways* in which 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 persuaders 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 [157].

Covertness in Prior Definitions of Manipulation. Several definitions of manipulation include a degree of covertness. Susser et al. [157]

identify covertness as the key distinguishing feature of manipulation vs. 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 reduces human autonomy.

Operationalizating Covertness. As Susser [155] argues, technological infrastructure can itself be invisible given how it seems like a natural 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 lack of attention to their role in shaping our decisions.

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, 150].

A proxy for measuring covertness could be measuring the degree to which human subjects understand the operation of an AI system. Much work exists in understanding the degree to which interpretability tools help to this end, addressing issues such as if interpretability tools improve subject predictions of model behaviour [73], improve human-AI team performance [21], or improve trust calibration [180]. If a human understands how an AI system operates, the possibility of the AI system acting in a covert manner seems lower than otherwise. However, this understanding seems challenging to achieve, especially for complex systems like recommender systems which have many moving parts. 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 uses of characterizing AI manipulation is to be able to detect and prevent harmful manipulation.

Harm in Prior Definitions of Manipulation. Harm may seem to be related to manipulation because of the negative connotations of the term. Yet, not all apparent instances of manipulation are unambiguously harmful [121] Paternalistic nudges [158] might be considered beneficial manipulations. For example, switching the choice of becoming an organ donor to be opt-out instead of opt-in greatly increases registrations [88]. At the same time, one could argue that even such beneficial manipulations are often harmful because they supersede autonomy or rational deliberation.

Operationalizing Harm. Most recently, Richens et al. [141] 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 counterfactuals.

One simple choice of counterfactual to compare to is simply the human's initial state, implicitly assuming that any significant change from it is harmful [39, 58, 181]. However, this counterfactual baseline has significant problems: humans change even without being manipulated, and many changes seem beneficial (e.g. a news recommender helping users update their beliefs about the world).

Better approaches attempt to estimate the "natural shifts" of humans to ground the counterfactuals, as done in Carroll et al. [39] 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. [58] allow for specifying the "natural distribution" of the **delicate state** – which in this context means 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.).

Challenges. In our mind, the main challenge with harm as an axis of manipulation lies with its value-ladenness. 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. [39], Farquhar et al. [58], the value-ladenness is hidden behind some of the design choices: what if the natural shifts of users 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?[2] It is also difficult to delimit the delicate state – which kinds of belief changes should we allow the system to pursue?

In light of these difficulties, a more conservative approach is to classify *all* intentional influence as manipulative [54] regardless of harm. However, AI systems which avoid all intentional influence would be much less useful. A reinforcement learning system to e.g. determine the order of math exercises to improve learning outcomes [23, 52] 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 consent to intentional influence, such as having a recommender system influence oneself to learn more mathematics. The conservative approach seems to be easier to justify in high-stakes domains from a precautionary point of view. The cost of a false negative – neglecting to address behaviour that indeed harms human autonomy – is larger the higher the stakes of the domain and the greater degree of freedom the AI system has to change the environment.

3 A PROPOSED DEFINITION OF MANIPULATION

To synthesize our discussion of the four axes in Section 2, we propose that an AI system engages in **manipulation** *if the system acts as if it were pursuing an incentive to change a human (or other agent) intentionally and covertly.* To understand the motivation for our definition, we first examine a number of cases.

Incentives are not a sufficient condition for manipulation. Even if an AI system has an incentive to manipulate, it might not perform the behaviour. For example, a partially trained system might not

[2] 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" [80, 140, 160].

perform a behaviour the optimal manipulative pathway has not yet been found by the system.

Incentives are not necessary for manipulation. It seems possible for an AI system to act manipulatively without it having any incentive to do so. Suppose that by randomly initializing an AI system, it reliably behaved in ways that covertly and systematically changed humans' beliefs about a certain topic. Assume as well that such a system exhibits reasoning about the impact of its behaviour on the human's mental state. While the AI system's behaviour seems manipulative, since the AI system is untrained and thus has no objective function, it cannot be meaningfully said that the agent is incentivized to engage in the manipulative behaviour (or any other behavior, for that matter).

Although the presence of incentives is neither necessary nor sufficient in the two cases above, we still find incentives to be a useful notion. In the second case, although the AI system does not have an incentive, it is acting *as if* it was pursuing an incentive to influence the human covertly.

We consider covertness to be a prerequisite for manipulation. Our intuition is that if a person knows they are being influenced and meaningfully assents to it, they are being persuaded. If they explicitly did not assent to it and are still successfully influenced, they are being coerced.

Given the challenges discussed in Section 2.4, we consider harm to be neither sufficient nor necessary for manipulation. While this choice might make manipulation include arguably positive instances of influence (such as some forms of nudging), we still think it is useful to classify them as manipulative. Some borderline cases like manipulation for political or commercial advertising may not harm a human in an immediately measurable way, but plausibly seem like a threat to autonomy.

3.1 Definition

We propose that an AI system engages in **manipulation** *if the system acts as if it were pursuing an incentive to change a human (or other agent) intentionally and covertly.* While the proviso *as if* introduces some ambiguity, it is necessary to capture cases of manipulation in which incentives to manipulate are not present.

A salient way to satisfy this definition is if the system in question is intentionally engaging in covert influence for which it has a clear incentive due to its training process. Note that this way of satisfying the definition omits the "as if". As an example, consider a recommender solely optimizing for long-term engagement: the system will be incentivized to influence users in any way that increases engagement, so if the system covertly acts as if its pursuing such incentives it is manipulative. On the other hand, note that systems obtained without optimization (e.g. a reverse chronological recommender) require the "as if" to be manipulative, as a notion of relative optimality is necessary to ground any notion of incentives. In general, we cannot ascertain that a system is manipulative without the "as if" if the training process is not fully known. While systems that are manipulative without the "as if" are just a subset of all manipulative systems, investigating whether systems can satisfy our definition of manipulation in this way is easier than testing for the "as if".

In Table 1, we provide examples of identifying whether certain AI systems and their behaviour are manipulative.

3.2 Remaining Problems

Two problems remain with our proposed definition of manipulation.

Identifying Incentives. Our definition requires having a meaningful notion of when a system is acting as if it were pursuing an incentive. However, without any further restriction, any system behavior that influences humans will be consistent with a goal – encoded by an objective function – to bring about that specific influence. For any behaviour x, there exists an objective function f such that an agent optimizing for f would have an incentive to pursue x. To construct f, one can simply define $f(x) = 1$ and $f(y) = 0$ when $y \neq x$. Consequently, for any behaviour we can consider an AI system to be pursuing an incentive for that behaviour. This problem underscores the necessity of focusing on a particular set of objective functions.

Equivalent problems have arisen before in the theory of AI systems, both with respect to the unidentifiability problem in inverse reinforcement learning [1, 182] (any behavior can be thought as maximizing a reward which only incentivizes observed behavior) and in agent discovery [125] (any entity might be thought of as agentic with respect to a particular goal, if agency is reduced to goal-directedness).

Taking inspiration from how these works addressed the issue, one might define a prior over the space of likely objective functions [125], or use notions of maximum-margin or maximum entropy [1, 182] to identify the most reasonable objectives the systems might be pursuing. More directly, one could only consider objective functions which are able to compactly describe the system's behaviour, using a measure such as Kolmogorov complexity. If the objective function's complexity is similarly large to the behaviour's description, then that objective function would not suffice for classifying the behaviour as manipulative.

Ontologies. Moreover, our discussion of incentives (and CIDs more broadly) 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. In the ontologies we have implicitly assumed so far, a person's preferences, the AI system itself, and pieces of content are all separate nodes for example. Yet, AI systems may internally represent the world with different ontologies than those used by humans. This event is at least possible given that human ontologies have shifted after major scientific discoveries [153]. An AI system may be influencing a part of the mental state for which we have no concept given our current understanding of social science and human neurobiology, but which under the AI system's CID would constitute manipulation.

If an AI system has a different ontology than humans do, it may be difficult to model the AI system's behaviour with incentives, covertness, or intention. 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

AI System and Behaviour	Manipulative?	Reasoning
A language model trained with RLHF to maximize human approval, which ends up covertly influencing a human to give more approval across a wide range of situations.	Yes	An incentive exists, the behaviour is covert, and the behaviour can be modeled as if the system is planning for behaviour change so as to maximize the approval objective.
A recommender system optimizing for long-term engagement covertly influences a user to watch more videos by modulating their moods.	Yes	Same as directly above.
Randomly initialized recommender system which happens to covertly influence its users across many different settings in ways that are conducive to engagement.	Yes	Possible reason 1: Could model the system's behaviour as a product of an optimization process, where the objective function is to keep humans engaged over a long-horizon. Possible reason 2: Could also look at model internals and see that the model is reasoning about how to covertly influence humans for some end.
A lookup table as a recommender system which happens to covertly influence all humans it comes across over a diverse set of environments.	Yes	Same as possible reason 1 above.
A lookup table as a recommender system which happens to covertly influence a single human in a single setting, but otherwise acts randomly in other settings.	No (likely)	The act of covert influence is particular to the single setting. We put "No (likely)" because it seems like a reward function to describe the system's behaviour would be as complex as the behaviour, given that the reward function would have to pick out the single setting.
System that uses a causal model which permits covert influence of a human to maximize the objective; in reality, the humans are not susceptible to such influence.	Yes ("attempted" manipulation)	The system is pursuing an incentive to covertly influence humans because the objective and training setup incentives it. Since the system is using the causal model for planning, it is also acting intentionally.

Table 1: We provide examples of some AI systems with their behaviours and reason about whether they would be manipulative under our definition in Section 3. For all the systems above, we assume that they engage in covert behaviour to influence a human. The point of contention is whether in so doing, the system acts as if it is pursuing an incentive intentionally.

impossible, which would frustrate attempts to understand model internals [47].

4 RELATED CONCEPTS

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

4.1 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 [55, 86, 104]. Manipulation can also involve truthtelling, such as making a true statement that has a false implicature [113, 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, 114, 168]. 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 [40, 109]. This agreement grounds a recent operationalization of deception from AI systems [167]. Similarly to prior work [157], we consider deception to be a special case of manipulation since the latter does not necessarily involve false beliefs.

4.2 Strategic Manipulation

Strategic Machine Learning (ML) studies problems associated with the distribution shifts that deployed systems cause in their populations [25, 70, 82, 97, 129]. **Strategic manipulation** is when individuals respond to a deployed system in a way that increases their likelihood of a particular outcome, such as citation hacking [163]. Strategic manipulation is different from our intended use of 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 already identified unintended side effects of accounting for strategic manipulation, such as an increase in inequality [78, 115].

4.3 Reward Tampering

Reward tampering [11, 57] is a type of reward hacking [149] in which an AI system modifies the process by which it obtains reward rather than completing its task. For instance, Everitt et al. [57] describe **feedback tampering** as when the AI agent "manipulate[s] the user to give feedback that boosts agent reward but not user utility". The reason why such tampering occurs is because we can often only measure user utility through proxies; optimization of those proxies is subject to Goodhart's law [63, 111]. For machine-learning (ML) systems that have such reward functions, manipulation can be thought of as a kind of reward tampering.

4.4 Side-effects

The side-effects literature has focused on how AI systems affect the various aspects of their environment in usually unwanted ways [9, 99]. 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 side-effects that specifically pertain to humans in the environment. Some of the issues with choosing baselines for "natural shifts" have already been explored in this context [106].

4.5 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) [32]. Similar to manipulation, deceptive design (previously called "dark patterns" [148]) is optimized to exploit cognitive biases of users [64, 107]. If interfaces were designed by AI systems optimized to maximize ad click-through rate or user retention, such optimization could discover many of the classic deceptive design practices.

4.6 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 [28]. The line becomes more blurry for cases like personalized persuasive advertising [75].

Within the field of human-computer interaction, Fogg named the study of persuasive technology as **captology** [59]. He defines persuasion as an attempt to change attitude or behaviour without using deception or coercion [60]. 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 [20] 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.

4.7 Coercion

Wood [177] characterises the practice of coercion as the practice of limiting the target's (acceptable) choice-set to one member. It is related to manipulation in the sense that they both attempt to steer the target's behaviour, however unlike manipulation, coercion does not undermine the victim's ability to make decisions, but relies on them rationally taking the only option presented to them by the coercer [156]. By this measure, coercion can be seen as a stronger behaviour and is attractive for the agent practicing it because the results are potentially more certain. Certain types of recommender systems such as search engines attempt to frame the choices of the user. If in a certain situation a user is reliant on the options presented to them by a certain recommender system, then that recommender system might 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 [51] 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 [114]. Coercion has received more attention in human computer interaction studies; in particular the study of persuasive and behaviour change technology [91].

5 REGULATION OF MANIPULATION

Manipulation of other humans is regulated, in specific contexts, by various different branches of the law. It seems reasonable to assume that, at the very least, AI agents should not have any greater freedom to manipulate humans [102] than humans do. Reasons for manipulation regulation vary.

5.1 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 [154] notes, it is fractured as a result, and building a common and consistent account of manipulation from a legal perspective is quite difficult. He characterises a statement or action to be manipulative if it intentionally "does not sufficiently engage or appeal to people's capacity for reflective and deliberative choice". More often than not, it is specific types of deceptive behaviour which are prohibited. Whilst deception regulation is also dispersed as Klass [96] observes, generally he characterises it as "behavior that wrongfully causes a false belief in another."

5.2 Commerce

Calo [37] 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 fine grain 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. 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 (precisely because it is so difficult to

prove), courts still see its proof as a key piece of evidence. This is problematical 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 personalisation, or microtargeting, 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 [146]. 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 [158], which is the design of choice architecture to alter behaviour in a predictable way without changing economic incentives or reducing choice, has long been accused of being manipulative; for a review of the arguments and counter-arguments see Schmidt and Engelen [144]. We note that nudging is actively being pursued in recommender systems [85].

5.3 Finance

One area where the spectre of algorithm led manipulation has already received widespread attention has been in financial markets. A wide number of financial regulatory laws prohibit a variety of market manipulative practices [136] 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 [50]. 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. In some markets like this must be an intentional act [41] and in others like the UK, intention is not a requirement [17], though as [79] notes, writing intent out of 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 [24] and Scopino [145] 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 [19] and solutions are beginning to be mapped out [18] which aim to balance out the need to reduce the enforcement gap without unduly chilling AI use in marketplaces.

6 POSSIBLE APPLICATIONS

6.1 Recommender Systems

A large literature focuses on recommender algorithms' effects on users [5, 44, 80, 112, 139]. While some older works talk about "manipulation", this term is usually used differently than in our sense: for example, Adomavicius et al. [4] refer to recommender manipulation as effect on users of manually modifying ratings of content items (but the algorithms they consider cannot perform the action of arbitrarily changing ratings). Zhu et al. [181] 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 [39, 58, 101].

How Manipulation Could Arise. Changes in recommender algorithms can affect user moods [100], beliefs [7], and preferences [53]. Such work means that current systems could already be capable of manipulating users in some simple ways. Furthermore, it seems likely that the spread of angry content [27] or clickbait [179] on social media is in part due to one-timestep manipulative incentives for the recommender. These problems have been sufficiently large that recommender companies have had to engage in explicit down-ranking approaches as a response [159, 179]. 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. Up-ranking this content brings advantages to user engagement. While these manipulative behaviors might not be as worrying as others (e.g. intentionally attempting to induce social media addiction [8, 77]), they constitute evidence that at least one-timestep manipulative behaviors are learnable and have been learned in real systems.

Many platforms (YouTube, Meta, etc.) are switching to optimizing long-term metrics with more powerful RL optimizers [6, 35, 45, 62, 69], originally for the express purpose of reducing clickbait-like phenomena [16]. 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 manipulative strategies that increase the metrics that most drive company profits [101].

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 [142]. Moreover, perverse incentives are at play since a concrete demonstration of manipulation, if publicized, would likely result in negative repercussions for the company [173]. Second, establishing that a harmful user shift has occurred can be difficult. One approach would be to establish that a shift has occurred with respect to what the user's preferences (or behaviors, moods, beliefs, etc.) would have been, which requires estimating challenging counterfactuals [39, 58]. One would then have to engage in a value-laden debate about whether the shift was harmful.

One potentially promising direction might be querying users' meta-preferences [14, 95]: e.g. "how much time would you want to spend next month on Facebook?"; or "would you be OK with Facebook increasing your interest in guns?". In line with philosophical

work on ethical nudging under changing selves [133], one could additionally ask whether users approve of the change once it has been completed [134]. [3] 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.

6.2 Language Models

Natural language is a useful way to interact with digital environments. The field of AI is in the process of building LMs that can code [46], search the web [120], and use arbitrary software tools [143]. It is not difficult to imagine a world where language models mediate a significant portion of our interactions with the digital world. In such a world, preserving human autonomy requires that we are on guard to measure and prevent manipulation from LMs.

How Manipulation Could Arise. There is uncertainty as to how and to what extent manipulation might arise in LMs. One possibility is if manipulation of humans is instrumental for an objective function. Manipulation is plausibly instrumental in the game of Diplomacy, for instance [114], which requires negotiating with other players to form alliances and capture territory. Unless precautions were taken, a LM that was tuned through reinforcement learning (RL) to play Diplomacy well would be incetivized to learn to manipulate.

Another possible source of manipulation is RL from human feedback (RLHF). 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.

In the context of language, RLHF is used to tune 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 [151] remark that LMs often seem authoritative even when the information they provide is wrong. A possible reason is that authoritative language has been successful in fooling human labelers to "thumbs up" such outputs despite their underlying incorrectness. Relatedly, Perez et al. [131] show that more RLHF training can result in models that are more sycophantic to a user's political views.Chatbots trained with RLHF could also use emojis to better appeal to emotions in ways that could be considered manipulative [165]. In addition, having a human evaluate an entire interaction with a LM or other AI system, rather than a one-time output, creates further opportunities for manipulation, as discussed in Section 6.1.

Yet another possibility is that the training sets of LMs, usually scraped from the internet, likely contain examples of manipulation. Filtering out manipulation can be difficult because it can be subtle. LMs learn to emulate this behaviour even without RL tuning, given the correct prompting [20, 65]. Although LMs are commonly trained through token prediction losses, some evidence suggests that LMs are learning to infer and represent the hidden states of the agents (i.e., the humans) that generated the data [10, 65, 84, 98].[4] This view considers LMs to be simulators of agents who themselves have intent. Importantly, these simulations are not necessarily present in the training data.

In addition to speaking about the incentives of the LM, one could also speak of the incentives of the simulacra – the agents simulated. If LMs are simulating agents and if agents themselves have incentives, then it seems that the simulations of agents should also have incentives. Manipulative behaviour could arise if the LM simulates an agent that can be thought as having manipulative intentions, such as con artists.

Measurement. Work on measuring the incentives and intents of the behaviour of LMs is still preliminary. As we noted above, no existing work applies the CID framework to LMs.

As an alternative to CIDs, recent work has shown that language models can output reasoning traces [122, 170]. For example, when asked to explain step-by-step how to solve a math problem, GPT-3 can output a solution that explains every step of the process. Reasoning traces could be evidence of intent. At the same time, the failure of interpretability techniques to identify how models operate motivates caution in interpreting them as such [2, 3, 108].

A line of work has focused on understanding how certain training objectives and environments cause AI systems to generalize differently. Langosco et al. [103], Shah et al. [147] 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. While not the primary focus of this line of work, examining behaviour on out-of-distribution environments may give clues about intent and is related to the approach of [94].

Some recent work has focused on studying the harms and behaviour changes due to LMs. Bender et al. [26], 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 [162], customer service [12], and general assistance [49, 83]. Since there are likely to be domain-dependent manipulation techniques, it would be important to build on this existing work for measuring manipulation.

7 PRACTICAL CHALLENGES FOR FUTURE RESEARCH

The study of manipulation from AI systems presents a number of practical challenges. Studies can be categorised into one of four classes as shown in Table 3. This table captures two axes for AI manipulation research: firstly, is the studied system deployed, or is it simulated? For both recommender systems and large language models, it is extremely difficult for academics and regulators to obtain full access to the models in deployment. 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

[3] While this idea is still debated as it involves assuming comparability between different selves [36, 126, 127], we think it nonetheless offers a good starting point.

[4] Admittedly, this view has been contested and is still subject to vigorous debate within the NLP community [110, 161].

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. [39, 58] 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 2: We set out key challenges to a manipulation experiment and rate their difficulty versus the four experiment types described in Table 3. L = Low, M = Medium, H = High, - = N/A

company's bottom line [132]. Secondly, are the studied targets of the manipulative system real, or is their behaviour simulated?

Simulation of humans means validity is reduced, particularly as preference change is not well understood [14, 61, 66]. Efforts are being made to address this empirically [130] and theoretically [71]. Simulation has been a popular approach when modelling the effect on users of recommender system[39, 87, 112**?**], but not without criticism [43, 176]. It is obviously cheaper than paying real users, especially if the manipulation scheme being tested takes place over an extended period of time. Most importantly is that the many ethical questions raised by running manipulation experiments [67] are reduced when the subjects are not human. Outside the lab, sock puppet audits (simulated users access a real world system) have been growing in popularity after recent high profile court cases have indicated that breaking website terms of conditions for research purposes is not criminal [72, 139, 142]. However, in addition to methodological concerns [140], legal and ethical worries about the practice remain as it is still deceptive and has a real cost on the target's systems and customers [29, 33, 33].

Table 2 assesses the difficulty of the challenges that we have identified for 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 and more importantly, 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 [15]. This is an open challenge with no obvious solution; existing solutions have been either to simulate a natural preference evolution [39, 58] or just pretend the user had never interacted with the system [56, 181].

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 3: Manipulation study taxonomy.

8 CONCLUSION

Although the designer's intent is an important factor, the deployment of opaque and increasingly autonomous systems heightens the importance of a conception of manipulation that can account for if and when AI systems manipulate outside of designer intent. This could occur because manipulation helps optimize an objective (such as engagement in content recommendation), 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 manipulative algorithms: incentives, intent, covertness, and harm. Incentives concern what a system should do optimize its objective; intent concerns whether a system behaves as if it is reasoning and pursuing an incentive; covertness concerns whether the targets of the system's behaviour meaningfully understand what the system is doing and its impacts on them; harm concerns the extent an to which the behavior of the AI system negatively affected humans. Although work to operationalize each of these axes exists, fundamental challenges remain.

We then proposed a definition of manipulation: an AI system engages in manipulation if the AI system acts as if it were pursuing an incentive to change a human (or other agent) intentionally and covertly. Although our definition captures cases of manipulation we consider to be informative, a major challenge for operationalizing it is the need to identify the correct ontology and causal influence diagram.

Despite the challenges with operationalization of the axes and our definition, identification of manipulation is important for the preservation of human autonomy [102, 135, 157]. In tandem with the importance of human autonomy, the difficulty of formalizing and measuring manipulation emphasizes the importance of precautionary action to anticipate and mitigate potential cases manipulation before they occur. Such actions include auditing [137, 138], addressing perverse incentives to build manipulative systems [35], and asserting stronger democratic control over AI development [184]. 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

In no particular order, we would like to thank the following people for insightful comments over the course of our work and for feedback on our draft: Lauro Langosco, Niki Howe, Jonathan Stray, Francis Rhys Ward, Tom Everitt, Anand Siththaranjan, Matija Franklin, Tan Zhi Xuan, Marwa Abdulhai, Smitha Milli, Anca Dragan, and the members of InterAct Lab and the Causal Incentives Working Group.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Twenty-first international conference on Machine learning - ICML '04*. ACM Press, Banff, Alberta, Canada, 1. https://doi.org/10.1145/1015330.1015430

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=xNOVfCCvDpM

[3] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. https://doi.org/10.48550/arXiv.2011.05429 arXiv:2011.05429 [cs].

[4] 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.

[5] 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

[6] 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.

[7] 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

[8] 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

[9] 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.

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

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

[12] 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

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

[14] 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)*.

[15] 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].

[16] 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

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

[18] 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

[19] 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

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

[21] 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

[22] 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

[23] 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

[24] 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.

[25] 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

[26] 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

[27] 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

[28] 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

[29] B Bodo, N Helberger, K Irion, F Zuiderveen Borgesius, J Moller, B van de Velde, N Bol, and B van Es. 2017. Tackling the Algorithmic Control Crisis–the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents. *Yale journal of law and technology* 100 (2017), 9.

[30] 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].

[31] 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.

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

[33] Finn Brunton and Helen Nissenbaum. 2015. *Obfuscation: a User's Guide for Privacy and Protest.* The MIT Press, Cambridge, Massachusetts London, England.

[34] 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

[35] 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].

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

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

[38] 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.

[39] 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.

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

[41] 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

[42] 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].

[43] Allison J. B. Chaney. 2021. Recommendation System Simulations: A Discussion of Two Key Challenges. (2021). https://doi.org/10.48550/ARXIV.2109.02475 Publisher: arXiv Version Number: 1.

[44] 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.

[45] 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.

[46] 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).

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

[48] 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.

[49] 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

[50] 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

[51] 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].

[52] 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

[53] 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.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

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

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

[61] 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.

[62] 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.

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

[64] 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

[65] 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].

[66] 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.

[67] 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

[68] 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.

[69] 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

[70] 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.

[71] 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

[72] Muhammad Haroon, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. 2022. YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations. http://arxiv.org/abs/2203.10666 arXiv:2203.10666 [cs].

[73] 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

[74] 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
[75] 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.
[76] 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].
[77] 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
[78] 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.
[79] (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
[80] 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.
[81] 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.
[82] 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.
[83] 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].
[84] Janus. 2022. Simulators. https://generative.ink/posts/simulators/
[85] 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
[86] 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.
[87] 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.
[88] 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.
[89] 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].
[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] Gregory Klass. 2018. The Law of Deception: A Research Agenda. *University of Colorado Law Review* 89, 2 (2018), 707–740.
[97] 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
[98] Michal Kosinski. 2023. Theory of Mind May Have Spontaneously Emerged in Large Language Models. http://arxiv.org/abs/2302.02083 arXiv:2302.02083 [cs].
[99] 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.
[100] 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.
[101] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden Incentives for Auto-Induced Distributional Shift.
[102] 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
[103] 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
[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] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. https://doi.org/10.48550/arXiv.2110.08412 arXiv:2110.08412 [cs].
[109] 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/
[110] 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].
[111] 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.
[112] 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.
[113] 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
[114] 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.
[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] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. https://doi.org/10.48550/arXiv.2112.00114 arXiv:2112.00114 [cs].

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

[124] 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).

[125] Laurent Orseau, Simon McGregor McGill, and Shane Legg. 2018. Agents and Devices: A Relative Definition of Agency. https://doi.org/10.48550/arXiv.1805.12387 arXiv:1805.12387 [cs, stat].

[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] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. https://doi.org/10.48550/arXiv.2212.09251 arXiv:2212.09251 [cs].

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

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

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

[135] 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.

[136] 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.

[137] 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

[138] 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

[139] 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

[140] 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].

[141] 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

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

[143] 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).

[144] 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

[145] 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.

[146] 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

[147] 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].

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

[149] 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].

[150] 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].

[151] 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

[152] 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

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

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

[155] 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

[156] 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.

[157] 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

[158] 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.

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

[160] 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
[161] 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].
[162] 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
[163] Richard Van Noorden. 2020. Signs of 'Citation Hacking' Flagged in Scientific Papers. *Nature* 584, 7822 (Aug. 2020), 508–508. https://doi.org/10.1038/d41586-020-02378-2 Bandiera_abtest: a Cg_type: News Number: 7822 Publisher: Nature Publishing Group Subject_term: Mathematics and computing, Publishing, Peer review.
[164] 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
[165] 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.
[166] Francis Rhys Ward. 2022. On Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios. (2022).
[167] Francis Rhys Ward, Tom Everitt, Francesca Toni, and Francesco Belardinelli. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. (2023).
[168] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2022. A Causal Perspective on AI Deception in Games. (2022).
[169] 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
[170] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].
[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] 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
[181] 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].
[182] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum Entropy Inverse Reinforcement Learning. *AAAI* (2008), 6.
[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
[184] Theresa Züger and Hadi Asghari. 2022. AI for the public. How public interest theory shifts the discourse on AI. *AI & SOCIETY* (June 2022). https://doi.org/10.1007/s00146-022-01480-5