SHAPE: A Framework for Evaluating the Ethicality of Influence

Elfia Bezou-Vrakatseli[1], Benedikt Brückner[2], and Luke Thorburn[1(✉)]

[1] King's College London, London, UK
{elfia.bezou_vrakatseli,luke.thorburn}@kcl.ac.uk
[2] Imperial College London, London, UK
b.brueckner21@imperial.ac.uk

Abstract. Agents often exert influence when interacting with humans and non-human agents. However, the ethical status of such influence is often unclear. In this paper, we present the SHAPE framework, which lists reasons why influence may be unethical. We draw on literature from descriptive and moral philosophy and connect it to machine learning to help guide ethical considerations when developing algorithms with potential influence. Lastly, we explore mechanisms for governing influential algorithmic systems, inspired by regulation in journalism, human subject research, and advertising.

Keywords: influence · manipulation · mental interference · nudging · choice architecture · suasion · persuasion · cognitive liberty · mental integrity · mental self-determination · freedom of thought · preference change

1 Introduction

Influence—which we define broadly as one agent taking an action that causes a change in another agent—is ubiquitous in multi-agent systems. If the agent being influenced is a person or is otherwise deserving of moral consideration, then it is widely accepted that some types of influence (e.g., blackmail, extortion) are unethical. In many settings where human communication is mediated by algorithms, however, the ethical status of influence is less clear. For example, interacting with a recommender system may change our preferences [25,43,64] and emotions [63], exposure to online political advertising can change our voting intentions [31], and interacting with large language models can change our opinions [10,58]. In such cases, it can be easier to sense that there may be an ethical principle being violated than to articulate the principle of concern.

There is a substantial body of work from descriptive and moral philosophy on concepts such as "influence" [102], "manipulation" [76], "mental interference" [37], "nudging" [91], "choice architecture" [94], "suasion" and "persuasion" [15], "cognitive liberty" [95], "mental integrity" [37], "mental self-determination" [21], freedom of thought [65], and preference change [25]. The definition of each of

V. Malvone and A. Murano (Eds.): EUMAS 2023, LNAI 14282, pp. 167–185, 2023.
https://doi.org/10.1007/978-3-031-43264-4_11

these terms, and the situations in which the phenomena they describe can be considered ethical, are all contested. We do not attempt to stipulate definitions or resolve normative disagreements in this paper. Rather, we draw on this literature to highlight specific reasons why some types of influence might be unethical, and link these concerns to relevant work from computer science and artificial intelligence (AI). Our hope is that the framework created through this synthesis will help designers of algorithmic systems that influence people to think more concretely about the ethical considerations relevant to their work.

Before introducing the framework we would like to stress that not all influence is bad or morally questionable. Our definition of influence (given in the opening sentence of this article) is so broad as to encompass all causal relationships between agents. In this view, all human communication—much of which is beneficial—constitutes a form of influence. In particular, rational persuasion ("the unforced force of the better argument" [52]) is often delineated as being a morally acceptable form of communication, and hence influence [98]. Without aiming to provide a perfect characterisation of wrongful influence, our view is that influence is ethically acceptable unless it possesses a property which makes it wrongful, and this paper is an attempt to compile a list of such properties.

Method. To arrive at the SHAPE framework we conducted an extensive (but unstructured) literature search in order to compile a list of reasons why influence may be unethical. This longlist of reasons was iteratively grouped into sets of similar concerns, discussed, supplemented with additional literature searches, and re-grouped until we arrived at the current version of the framework. This process was not straightforward due to the fact that a number of categories have a non-negligible overlap, and our decisions about how to hierarchically arrange the relevant ideas are inevitably somewhat contingent and subjective. Additionally, we emphasise that this article is not a true systematic review, and the amount of literature relevant to the ethics of influence is vast. Nonetheless, we are confident that the chosen categories are informative, if not perfectly disjoint, and to the best of our knowledge we are the first to provide a framework to assess the ethicality of influential AI systems.

2 Concerns

In this section we develop our SHAPE (Secrecy, Harm, Agency, Privacy, Exogeneity) framework by listing reasons why influence may be unethical, drawing on work from moral philosophy and linking it to relevant concepts in computer science and AI. We do not claim that this list is comprehensive, but we do think it covers the most commonly-cited objections to influence. Similarly, we do not claim that this is a perfect taxonomy or that each of the reasons given is perfectly distinct from the others, but we do argue that the five groups of reasons—secrecy, harm, agency, privacy, and exogeneity—capture meaningful families of objections. The aim of the framework is to provide guidance about whether a particular instance of influence might be unethical to those in charge of designing the agent or process which exerts the influence. While the terms in

The SHAPE framework

Influence might be unethical if...

Secrecy (2.1)

... the influencee is not fully aware of the intent to influence
... the influencee is not fully aware of the means of influence

Harm (2.2)

... it causes harm

Agency (2.3)

... it removes options
... it imposes conditional costs or offers
... it occurs without consent
... it bypasses reason
... it is not able to be resisted

Privacy (2.4)

... it breaches an assumed contract regarding the use of personal information
... it is implicated in mass surveillance

Exogeneity (2.5)

... it serves exogenous rather than endogenous interests
... it gives one group power over another

Box 1. The SHAPE framework for considering the ethicality of influence.

the acronym give an overview of concerns, the corresponding sections provide a more detailed analysis for each. An overview of the framework summarising these reasons is given in Box 1, and a discussion of the concept of intent (which is relevant to all five reasons) is given in Box 2.

2.1 Secrecy

First, influence may be unethical if it involves *secrecy*. In the literature, variations of this idea have also been referred to as "covertness" [39], "deception" [72], "lying" [71], or "trickery" [76]. Articulating precise definitions for these terms is an open philosophical problem [67], but many have been proposed. The core idea is perhaps most neutrally defined as an "information asymmetry", where the influencer has more information than the influencee [35]. More narrowly,

deception has been defined as any situation where an agent A intentionally causes another agent B to have a false belief, with necessary requirement that agent A does not believe it to be true [26].

Secrecy of all sorts may be wrong—when it is wrong—because it violates a moral norm or duty, specifically "a duty to take care not to cause another to form false beliefs based on one's behaviour, communication, or omission" [97], because it constitutes a breach of an implicit promise to be open and truthful [84], or because it constitutes a betrayal of trust [71]. The wrongness of secrecy may also in some cases be due to downstream consequences of the secrecy, rather than due to the secrecy itself. For example, some argue that when an intent to influence is hidden from the influencee, it is "less likely to trigger rational scrutiny" [76] and thus bypasses reason, reducing agency (Sect. 2.3).

That said, secrecy may not always be unethical, as in cases of "benevolent deception" [3]. For example, it may be beneficial for the rehabilitation of patients who have suffered stokes or other brain injuries if their physical therapist robot obfuscates their true progress towards recovery [19].

Here, we distinguish between two types of secrecy as it relates influence: secrecy of *intent* and secrecy of *means*.

Secrecy of Intent. Influence may be unethical if the influence is intended by the influencer, and the influencee is not fully aware of this intent. For example, a video deepfake [73] intended to influence public opinion in a certain direction (perhaps by misrepresenting the actions of a political figure) may be unethical because the people who are influenced are not made fully aware of this intention. Had they been aware, they would have assigned less credence to the information contained in the video [68].

Secrecy of Means. Influence may also be unethical if the influencee is not fully aware of the means by which they are being influenced. For example, a user interacting with a sophisticated social media recommender system may be fully aware that the algorithm is designed to maximise the total amount of time they spend on the platform—so there is no secrecy of intent—but be unaware of the strategies the recommender is employing to achieve this, such as through the occasional recommendation of content that is increasingly sympathetic to a conspiracy theory [105].

Technical Work. Of the many ethical objections to influence, secrecy has perhaps received the most attention in the context of AI. For example, the sizable literature on algorithmic transparency, explainability, and interpretability (see, e.g., [27,69]) represents an attempt to mitigate information asymmetries between AI systems and their human users. There is also an emerging literature that seeks to provide formal definitions of deception from a causal perspective, along with mechanisms for detecting it in AI systems [88,113,114,116]. Algorithmic agents can also fall prey to influence involving secrecy, as in cases of adversarial attacks [70], data-poisoning [70], reward function tampering [44], and manipulating human feedback [115].

2.2 Harm

Second, influence may be unethical if it causes *harm*. There are many different forms of harm, with some of the most prominent categories including reduced physical or mental well-being [78], bias [118], unfairness [118], or injustice [101]. In general, harm and related concepts such as "suffering" [56] are expansively but inconsistently defined. Definitions range from those that equate harm with any "physical or other injury or damage" [23], to those state harm is a condition of "interference with individual liberty", originating from the "harm principle" of John Stuart Mill [81], a definition which would liken harm to a reduction in agency (Sect. 2.3).

Ethical (if not legal) views on what does and doesn't count as harm are normative and contested, and this is notably true of harms that may arise from speech acts in algorithmically-mediated online fora. For example, "safe spaces" are viewed by some as a means of avoiding psychological harm and others as an institution which, if realised, inflicts epistemic harms [6]. Regardless of the position one takes in such debates, it seems defensible that there are many forms of harm which are widespread but not frequently well-articulated, and some of these harms can plausibly be promulgated by influential AI systems. One example of such harms has been labelled epistemic injustice [61]. Varieties of epistemic injustice include *testimonial injustice*, where an individual is discredited as a credible source of knowledge, and *hermeneutic injustice*, where an individual experiences reduced capacity to make sense of their own experiences due to a lack of a relevant framework, shared vocabulary, or common knowledge of a shared experience. Both forms of epistemic injustice may be exacerbated by language models or recommender systems, if such systems are heavily used and systematically privilege certain perspectives.

It should be emphasised that harm, while perhaps intrinsically injurious, need not always be unethical. A surgeon making a cut to a patient's skin to fix their broken leg may cause temporary harm and pain, but is arguably acting in the best interests of the patient. In such cases, influence would then not be unethical despite causing harm. The assessment of harmful influence is further complicated by the fact that it can be very hard to define when influence is actually harmful, particularly influence over mental properties such as preferences [25].

Technical Work. The concept of harm is a central topic among AI policymakers, with the prevention of harm being underscored as a critical principle for AI systems in the European Commission's report on trustworthy AI. The report asserts that AI systems should never cause adverse effects on any human being [54]. Harm, particularly in the physical sense induced by AI systems, has been scrutinized extensively within the domain of self-driving cars through thought experiments like the trolley problem [41].

Another significant area of research is AI in healthcare, where there is a strong emphasis on the minimization of harm potential. AI systems in healthcare are expected not only to elevate the well-being of individuals but also to consider the

Intent

When debating ethical considerations concerning AI systems, the concept of intent is highly relevant. The definition we have adopted for deception already encompasses the notion of intention, but it is easy to see that, for example, whether harm that was caused or a reduction of agency that took place was intended by a given agent also requires a more thorough definition of this notion. While the concept of intention may be somewhat easy to understand for humans, it becomes more obscure as soon as algorithmic agents are involved. Such agents may deploy deceptive, harmful or otherwise undesired strategies without human intention [65]. A key distinction is, thus, to be made between reprehensible actions that follow human intention and such actions that happen unintentionally (regarding the human responsible). In the case of deception in human-machine relations three distinct cases are: (1) the agent deceives as a result of human intention to deceive; (2) the agent deceives autonomously or incidentally without human intention; and (3) the agent deceives its designer [31].

Technical Work Intent in AI systems that deceive or manipulate is analysed in a number of works. It can be seen as a key dimension of manipulation, but may be hard to define and operationalise [27,62]. Helpful definitions of intent for algorithms may be derived based on notions from legal theory since parallels between judging whether a human intended to commit a crime and judging whether an artificial agent intended to perform an action such as deceiving its creator or causing harm to a user can be drawn [8]. However, even if suitable definitions are found, intent and causation tests may fail for black-box AI algorithms, implying that different approaches to the issue may be necessary [14].

Box 2. The concept of *intent*, as it relates to the ethics of influence.

potential psychological or mental harm they may cause, such as those resulting from discrimination or neglect [78].

A prominent challenge in this field is assigning responsibility when harm does occur, given the numerous actors typically involved in the development process. This issue is particularly salient in the context of recommender systems, which often serve to influence human behaviour. Even when these systems are designed with benevolent intentions—such as supporting healthy decision-making—they can unintentionally cause adverse effects [40].

2.3 Agency

Third, influence may be unethical if it reduces human *agency*, or related concepts such as "self-determination" [21] and "autonomy" [85]. There are many proposed definitions of agency [46]. One account defines agency as the act of an agent making use of its ability to act [90]. In this view, agency requires that executed actions are intended, and result in part from the agent's reasoning processes. To

reduce human agency, then, is to disrupt the link between an agent's intentions or reasoning processes and their subsequent actions.

Several works link influence with a reduction in agency. Being influenced into performing an action reduces the agency of an individual, at least in terms of the decision about whether to perform that action [103]. Human agency is often characterised as having intrinsic moral value, and reductions in agency may be wrong regardless of whether that reduction in agency is paternalistic and results in improved welfare for the person affected. Not respecting the competency of an individual to make their own decisions is seen as a lack of appreciation of them being a rational agent [96] or even a degradation [75]. Perhaps more unambiguously, reduced agency can be wrong if it involves impairments to the psychological capabilities of the subject thought to be the basis for free will [100]. The wrongness of reductions in human agency may also stem from the fact that the interests of the affected agent are being devalued or deprioritised relative to those of the another party (see Sect. 2.5) [86,96].

However, it has also been argued that reductions in agency are not always wrong, and that rational agents often do not oppose influence that has this effect [22]. Instead, agency may be valuable instrumentally because is often a useful means to an end. We sometimes place ourselves in situations where we have reduced agency—such as following a recipe or studying a prescribed curriculum—if it helps to achieve a goal.

Here, we give five accounts of what it means for influence to reduce agency: removing options, imposing conditional costs or offers, influencing without consent, bypassing reason, or being irresistible. These are likely not mutually exclusive.

Removal of Options. Influence may be unethical if it removes options previously available to the influencee [49]. For example, an autonomous vehicle may in some implementations prevent its human driver from deciding to take a certain route to a destination that they otherwise would have taken. Options may be removed explicitly (by refusal) or implicitly (by a failure to provide an affordance that would enable the option). Options can also be removed effectively, without being absolutely removed, by imposing conditional costs (see below) that are so severe as to make the option untenable. Such removal of options, where the influenced party can be said to have no choice or no acceptable choice, has been labelled "coercion" [62,77,119].

Conditional Costs or Offers. Influence may be unethical if it imposes conditional costs or offers on the influenced depending on the action they choose to take, thus altering the relative appeal of different options. In philosophical literature, this type of influence is sometimes called "pressure" [76]. Conditional costs can be seen as a form of threat, though the severity of the threatened cost can vary significantly. Examples of costs that might be threatened include a loss of time or energy (e.g., nudging [101] or browbeating [12]), a loss of social status (e.g., peer pressure), or physical violence (e.g., kidnappers demanding a ransom).

It is possible to use carrots as well as sticks: the costs imposed may be opportunity costs. For example, the influencer may attach positive incentives or "offers" (e.g., money or status) to certain alternatives, which reduces the relative value of others [87]. Such incentives are not always unethical. For example, it is generally considered acceptable to offer salaries to influence people to work for you. Baron [12] suggests that such incentives are only unethical if they mean the influenced adopts a particular alternative for "the wrong sort of reason" [12]. Which sorts of reasons are considered wrong will be context specific.

Consent. Influence may be unethical if it occurs without (informed) consent, thus potentially ignoring a decision a person has made while exercising their agency [45]. For example, consent is plausibly the morally distinguishing factor between strenuous exercise and forced labour.

Bypassing Reason. Influence may be unethical if it bypasses human reason [51]. Mechanisms of influence which involve the bypassing of reason include: customised presentation of information, the flooding of agents with irrelevant information to crowd out relevant information, and the withholding of certain information [17]; exploitation of known imperfections in human decision-making such as group pressure [7]; exploitation of the "truth effect", which is the fact that frequent repetition of a statement increases the probability of individuals to find that statement to be true [53,92]; anchoring [5]; and appeals to emotion such as fear [57].

Irresistibility. Influence may be unethical to the extent that it is difficult to resist [17,28]. Attempts at influence can be made difficult to resist through the use of techniques such as flattery or seduction. Use of such techniques arguably reduces agency of those influenced. This has direct implications on the moral responsibility of an agent for their actions. Such responsibility has been claimed to not require "regulative control", i.e. access to alternative possibilities, but merely "guidance control" as control over the mechanism which steers their behaviour. An agent who is influenced into acting in a certain way through mechanisms they cannot resist is therefore not morally responsible for the consequences of their actions [47].

Technical Work. There is an emerging body of technical work that seeks to quantify degrees of agency, often from a causal perspective [29,60]. There has also been work that seeks to use AI to support human agency in certain contexts, such as in learning environments [34] or on social media platforms [59].

2.4 Privacy

Influence may also be unethical if it is made possible by a violation of *privacy.* Privacy is a fundamental aspect of our lives that refers to our ability to control access to our personal information. It encompasses the right to keep certain

information about ourselves hidden from others and is vital for protecting our individuality, fostering trust, and preserving our personal freedom. The more information is known about a person, the greater the extent to which it is possible to identify mechanisms by which they can be influenced. Nissenbaum [74] identifies three privacy principles frequently cited when justifying privacy-enhancing laws: (1) limiting surveillance of citizens and use of information about them by agents of government, (2) restricting access to sensitive, personal, or private information, (3) curtailing intrusions into places deemed private or personal.

In the first years after the internet was established a number of very serious invasions of individual privacy were committed [110]. There is currently a consensus on condemning such actions, but the concern of privacy is still relevant and a very complex one. When training an agent, privacy can be inadvertently breached through data collection, data aggregation, predictions or third-party access [120]. One example of a practice that often raises privacy concerns is personalised ads. The extensive collection of user data raises concerns about the transparency of data collection practices and the potential for unauthorised access or misuse of personal information [109]. More generally, the personalised, virtual experience that such practices result in "fractures the public sphere into individual parallel realities" [110], while also being more likely to promote extreme content, and less likely to be noticed by experts who have historically been responsible for fact-checking (e.g., journalists).

As the concern of privacy is very complex, it is important to be able to identify the type of information that is private and which should therefore be protected (and not used without our consent). Ben and Lazar [13] distinguish between the following types of data: *training* (i.e., data collected to train predictive models) vs *targeting* (i.e., data used for targeting); *sensitive*[1] (i.e., data about a person that they might reasonably not want others to know) vs *nonsensitive*; and subdivide *sensitive* into *intrinsically sensitive*(i.e., if it is sensitive when considered on its own) vs *extrinsically sensitive* (if it is sensitive only when considered in combination with other data points). Privacy concerns arise when the training data consists of sensitive and nonsensitive information [11]; a model trained on that data can uncover a link between intrinsically nonsensitive properties P, Q, and R, and intrinsically sensitive property S. This means that if we have access to values for these non-sensitive properties for a user, the chances of successfully predicting S increase [13].

We address the privacy concern on two levels: as an individual breach of contract or trust, and as a wrong associated with collective surveillance.

Breach of Contract. Thinking back to the three privacy principles, principles (2) and (3) address the individual level. A privacy breach constitutes a violation of these principles. Principle (3) encompasses the traditional idea of *sanctity*, in support of the notion of people "shielding themselves from the gaze of others", whereas principle (2) encapsulates the nature of the information collected, and

[1] An extensive analysis of the notion of "sensitive information" and why it is critical can be found in [111].

potentially disseminated, which should be protected when it meets societal standards of intimacy, sensitivity, or confidentiality [74]. The ethical ramifications of influence encompass the broader societal implications of privacy violations; a breach of contract in these cases constitutes a degradation of human dignity. This extends beyond the individual level since individual privacy infringements can violate the right to privacy of other people, and the consequences of privacy losses are experienced collectively [110].

Surveillance. The issue of surveillance adds an extra layer to the aforementioned collective experience of privacy loss. The first of the three principles is dedicated to this concern, and it constitutes a special case of the more general principle of protecting individuals against unacceptable government domination. The right to privacy can thus also be understood by referring to general, well-defined, and generally accepted political principles addressing the balance of power [74] (See also Sect. 2.5). An invasion of the privacy of an agent gives others power over that agent [110]. On a societal level, citizens' autonomy is threatened when they lose their privacy. The more data are collected, the easier it becomes to anticipate the following actions of an individual, the more prone people become to influence, and the easier it becomes to justify this influence. Government surveillance becomes, thus, more powerful once they gain access to said data. This is a critical concern since "a largely unregulated tech industry is detrimental to free and democratic societies" [110].

Technical Work. While the most obvious approaches to mitigating privacy concerns relating to influence involve simply deciding whether or not to proceed with a given product deployment or research project, there is also research on technical approaches to respecting privacy in certain applications of influential AI. These include work on differential privacy [1,38] and contextual integrity [14,33].

2.5 Exogeneity

Lastly, influence may be unethical if it advances interests not held by the agent being influenced, a property we call *exogeneity*.

We present two articulations of unethical exogeneity in influence: the disparate advancing of exogenous and endogenous interests, and the exercise of power.

Exogenous Interests. Influence may be unethical if it advances exogenous goals or interests (those not held by the influencee) over endogenous goals or interests (those held by the influencee). In this account, the wrongness of influence stems not from the fact that the influencer benefits (they may not benefit), or from harm to the influencee in absolute terms (they may not be harmed), but from the relative advantaging of the interests of another agent over the interests of the influencee [13,76,86].

Power. Influence may also be unethical if it empowers one party over another, or constitutes an exercise of power of one party over another. There is considerable philosophical literature on how power is instantiated in technology [16], as well as related concepts including "control" and "domination" [9]. For example, manipulating the opinion of a single individual can be difficult [31], but widely-used recommender systems present a vector by which a minority might steer the opinions and behaviour of a larger population, through an accumulation of small or stochastic effects. Another example of power being abused is the use of AI-enabled ad targeting to influence election results [18].

Technical Work. Monitoring whose interests are being served through the use of an AI system lends itself naturally to questions of fairness, and there is substantial literature on both formal measures of fairness [82] and algorithms for promoting it [112]. Another relevant line of work relates the development of mechanisms for diffusing or decentralising the power that is exercised through the use of influential AI systems. This includes both technical social choice mechanisms for choosing objective functions [66], and the use of participatory institutions such as citizen assemblies [79] and collective response systems [32,80] to provide democratic oversight.

3 Governance of Influence

For the most part, the concerns listed in Sect. 2 point to general or abstract principles that can inform an understanding of the ethical status of different kinds of influence. In order for such an understanding to be widely adopted into the practices of those designing and building influential algorithmic systems, we need mechanisms for deciding, disseminating and enforcing what best practice looks like in specific, concrete terms. Here we point to three such mechanisms (professional cultures, ethics review processes, government regulation) via examples from other domains (scientific research with human subjects, journalism, advertising).

3.1 Professional Culture

In journalism there is minimal formal oversight of ethical practice, but nonetheless there is broad understanding of a core set of ethical principles which are reinforced by educational institutions, professional organisations, and workplace culture [48,89]. These principles commonly include mention of accuracy or truthfulness [83], objectivity or impartiality [117], and avoidance of harm through the use of anonymity or avoiding coverage of certain topics (e.g. suicide) [24,36]. Such principles informally govern influence in the context of journalism. Similar ethical principles exist in computer science, but these are not as widely adopted [20,30].

3.2 Institutional Ethics Reviews

Formal ethics review processes, such as those conducted by most academic institutions in advance of research that involves human subjects, are one way of formalising a consideration for the ethics of influence. Reviewers involved in such processes already grapple with the use of techniques such as deception or trickery to create experimental conditions [8], and with what it means to have meaningfully consented to be subject to such influence [55]. Examples of such review processes in practice are numerous, in AI research a number of prestigious conferences and journals have implemented such mechanisms through checklists and the provision of guidelines [99]. The same holds true for industry where the widespread deployment of AI-based algorithms has lead to the establishment of ethics review processes by large companies such as Adobe or Google [4,50].

3.3 Regulation

In many jurisdictions, the advertising industry is subject to laws that place limits on the content of advertising and the contexts in which certain types of advertising can appear. These often require that advertising avoid outright deception (e.g., truth-in-advertising laws) [106], and ban ads in contexts where they are thought to cause harm (e.g., the ban of gambling, alcohol, or fast food ads during childrens' programs or televised sports) [2,104]. Such laws formally specify classes of influence which are collectively deemed unacceptable in the context of advertising.

Since AI is a fast-moving field, implementing regulatory guidelines for it presents a challenge. Though not specifically targeted at AI systems, the European Union's General Data Protection Regulation (GDPR) sets out a number of rules which implicitly impose constraints on Artificial Intelligence as well [93]. These rules will be concretised by the Union's Artificial Intelligence Act which it aims to pass by the end of 2023 and which is specifically targeted at the regulation of AI Systems [42]. Further examples of planned AI regulation include the attempts in the United Kingdom where a white paper was recently published which will be used as the basis for the country's AI regulations [107] as well as the US which published a Blueprint for an AI Bill of Rights [108].

4 Conclusion

In this paper we have synthesised some of the most commonly cited reasons—captured by the acronym SHAPE—why influence can be unethical. Specifically, these are that influence can (1) involve *secrecy* regarding the intent or means of influence, (2) cause *harm*, (3) reduce human *agency* by removing options, imposing conditional costs or offers, occurring without consent, bypassing reason, or being irresistible, (4) violate *privacy* by relying on the use of private information in a way that breaches an assumed contract or being implicated in mass surveillance, and (5) advance *exogenous* interests at the expense of endogenous interests, or give one group power over another. We linked each of these

general principles to relevant concepts from computer science and artificial intelligence, and described three models of ethical governance from other domains—professional culture which emphasises ethics, institutional ethics reviews, and regulation—which could be employed to translate such general principles into practice.

We envisage the SHAPE framework being used by designers of influential AI systems as a way to structure their thinking when considering the ethical impacts of their systems. For example, those building a product based on a large language model (LLM) might systematically work through Box 1, enumerating the examples of each of the SHAPE concerns that arise in the context of their product. These might include user-to-LLM feedback loops that are not understood by the user (*secrecy*), defamatory hallucinations (*harm*), affordances that require extra effort by users to surface certain perspectives in model outputs (*agency*), use of personal data to improve user retention (*privacy*), and adversely paternalistic choices in the design of the product (*exogeneity*), among others. Such a list could then be translated into a list of actions to be taken to remove or mitigate each of these ethical concerns.

For the most part, we have in this paper refrained from stipulating particular definitions or drawing definitive lines between ethical and unethical influence. Such decisions will likely be context-specific and contested, and our focus has instead been on drawing connections between work in philosophy and computer science. That said, it would be valuable for future work to consider the extent to which these concerns over influence could be made more precise by focusing on narrower domains, such as LLM-enabled chat interfaces or social media recommender systems.

Acknowledgements. The authors were supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (safeandtrustedai.org), co-located at King's College London and Imperial College London.

References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318. CCS 2016, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2976749.2978318
2. Adams, J., Tyrrell, R., Adamson, A.J., White, M.: Effect of restrictions on television food advertising to children on exposure to advertisements for 'less healthy' foods: Repeat cross-sectional study. PLOS ONE **7**(2), 1–6 (2012). https://doi.org/10.1371/journal.pone.0031578
3. Adar, E., Tan, D.S., Teevan, J.: Benevolent deception in human computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1863–1872 (2013)
4. Adobe Inc.: AI Ethics. https://www.adobe.com/uk/about-adobe/aiethics.html
5. Adomavicius, G., Bockstedt, J.C., Curley, S.P., Zhang, J.: Do recommender systems manipulate consumer preferences? A study of anchoring effects. Inf. Syst. Res. **24**(4), 956–975 (2013). https://doi.org/10.1287/isre.2013.0497

6. Anderson, D.: An epistemological conception of safe spaces. Soc. Epistemology **35**(3), 285–311 (2021). https://doi.org/10.1080/02691728.2020.1855485
7. Asch, S.E.: Opinions and social pressure. Sci. Am. **193**(5), 31–35 (1955). https://doi.org/10.1038/scientificamerican1155-31
8. Athanassoulis, N., Wilson, J.: When is deception in research ethical? Clin. Ethics **4**(1), 44–49 (2009). https://doi.org/10.1258/ce.2008.008047
9. Aytac, U.: Digital domination: Social media and contestatory democracy. Polit. Stud. 00323217221096564 (2022). https://doi.org/10.1177/00323217221096564
10. Bai, H., Voelkel, J.G., Eichstaedt, J.C., Willer, R.: Artificial intelligence can persuade humans on political issues (2023). https://doi.org/10.31219/osf.io/stakv. https://osf.io/stakv/
11. Barocas, S., Nissenbaum, H.: Big data's end run around anonymity and consent. Priv. Big Data Public Good: Frameworks Engagem. **1**, 44–75 (2014)
12. Baron, M.: Manipulativeness. In: Proceedings and Addresses of the American Philosophical Association, vol. 77, no. 2, pp. 37–54 (2003). http://www.jstor.org/stable/3219740
13. Benn, C., Lazar, S.: What's wrong with automated influence. Can. J. Philos. **52**(1), 125–148 (2022). https://doi.org/10.1017/can.2021.23
14. Benthall, S., Gürses, S., Nissenbaum, H., et al.: Contextual integrity through the lens of computer science. Now Publishers (2017)
15. Berdichevsky, D., Neuenschwander, E.: Toward an ethics of persuasive technology. Commun. ACM **42**(5), 51–58 (1999). https://doi.org/10.1145/301353.301410
16. Bloomfield, B.P., Coombs, R.: Information technology, control and power: the centralization and decentralization debate revisited. J. Manage. Stud. **29**(4), 459–459 (1992)
17. Blumenthal-Barby, J.S.: A framework for assessing the moral status of manipulation, In: Weber, C.C.M. (ed.) Manipulation, pp. 121–134. Oxford University Press (2014)
18. Boine, C.: AI-enabled manipulation and EU law (2021). https://doi.org/10.2139/ssrn.4042321
19. Brewer, B.R., Fagan, M., Klatzky, R.L., Matsuoka, Y.: Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. IEEE Trans. Neural Syst. Rehab. Eng. **13**(1), 1–11 (2005)
20. BCS, The Chartered Institute for IT: Code of conduct for BCS members (2022). https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf
21. Bublitz, J.C., Merkel, R.: Crimes against minds: on mental manipulations, harms and a human right to mental self-determination. Crim. Law Philos. **8**(1), 51–77 (2014). https://doi.org/10.1007/s11572-012-9172-y
22. Buss, S.: Valuing autonomy and respecting persons: manipulation, seduction, and the basis of moral constraints. Ethics **115**(2), 195–235 (2005). https://doi.org/10.1086/426304
23. Cambridge dictionary (2023). https://dictionary.cambridge.org. Accessed 23 July 2023
24. Carlson, M.: Whither anonymity? journalism and unnamed sources in a changing media environment. In: Journalists, Sources, and Credibility, pp. 49–60. Routledge (2010)
25. Carroll, M., Hadfield-Menell, D., Russell, S., Dragan, A.: Estimating and penalizing preference shift in recommender systems. In: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 661–667. RecSys 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460231.3478849

26. Carson, T.L.: Lying and Deception: Theory and practice. OUP Oxford, Oxford (2010)
27. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832. https://www.mdpi.com/2079-9292/8/8/832
28. Cave, E.M.: What's wrong with motive manipulation? Ethical Theor. Moral Pract. **10**(2), 129–144 (2007). https://doi.org/10.1007/s10677-006-9052-4
29. Chan, A., et al.: Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 651–666. FAccT 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594033
30. Association for Computer Machinery: ACM code of ethics and professional conduct (2018). https://www.acm.org/code-of-ethics
31. Coppock, A., Hill, S.J., Vavreck, L.: The small effects of political advertising are small regardless of context, message, sender, or receiver: evidence from 59 real-time randomized experiments. Sci. Adv. **6**(36), eabc4046 (2020). https://doi.org/10.1126/sciadv.abc4046
32. Coy, P.: Can A.I. and democracy fix each other? New York Times (2023). https://www.nytimes.com/2023/04/05/opinion/artificial-intelligence-democracy-chatgpt.html
33. Criado, N., Such, J.M.: Implicit contextual integrity in online social networks. Infor. Sci. **325**, 48–69 (2015). https://doi.org/10.1016/j.ins.2015.07.013
34. Deschênes, M.: Recommender systems to support learners' agency in a learning context: a systematic review. Int. J. Educ. Technol. High. Educ. **17**(1), 50 (2020). https://doi.org/10.1186/s41239-020-00219-w
35. Dierkens, N.: Information asymmetry and equity issues. J. Financ. Quant. Anal. **26**(2), 181–199 (1991)
36. Domaradzki, J.: The Werther effect, the Papageno effect or no effect? A literature review. Int. J. Environ. Res. Public Health **18**(5), 2396 (2021). https://doi.org/10.3390/ijerph18052396
37. Douglas, T., Forsberg, L.: Three rationales for a legal right to mental integrity. In: Ligthart, S., van Toor, D., Kooijmans, T., Douglas, T., Meynen, G. (eds.) Neurolaw. Palgrave Studies in Law. Neuroscience, and Human Behavior, pp. 179–201. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69277-3_8
38. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
39. Dynel, M.: Comparing and combining covert and overt untruthfulness: on lying, deception, irony and metaphor. Pragmatics Cogn. **23**(1), 174–208 (2016)
40. Ekstrand, J.D., Ekstrand, M.D.: First do no harm: considering and minimizing harm in recommender systems designed for engendering health. In: Engendering Health Workshop at the RecSys 2016 Conference, pp. 1–2. ACM (2016)
41. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Ethics **21**(4), 403–418 (2017). https://doi.org/10.1007/s10892-017-9252-2
42. European Parliament: EU AI Act: First regulation on Artificial Intelligence (2023). https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
43. Evans, C., Kasirzadeh, A.: User tampering in reinforcement learning recommender systems (2022)

44. Everitt, T., Hutter, M., Kumar, R., Krakovna, V.: Reward tampering problems and solutions in reinforcement learning: a causal influence diagram perspective. Synthese **198**(Suppl 27), 6435–6467 (2021)
45. Faden, R.R., Beauchamp, T.L.: A History and Theory of Informed Consent. Oxford University Press, Oxford (1986)
46. Ferrero, L.: An introduction to the philosophy of agency. In: The Routledge Handbook of Philosophy of Agency. Routledge (2022)
47. Fischer, J.M.: Responsibility and manipulation. J. Ethics **8**(2), 145–177 (2004). https://doi.org/10.1023/B:JOET.0000018773.97209.84
48. Frost, C.: Journalism Ethics and Regulation. Taylor & Francis, Milton Park (2015). https://books.google.co.uk/books?id=K5b4CgAAQBAJ
49. Garnett, M.: Agency and inner freedom. Noûs **51**(1), 3–23 (2017). http://www.jstor.org/stable/26631435
50. Google LLC: Google AI Review Process. https://ai.google/responsibility/ai-governance-operations/
51. Gorin, M.: Do manipulators always threaten rationality? Am. Philos. Q. **51**(1), 51–61 (2014)
52. Habermas, J.: Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. The MIT Press, Cambridge (1996)
53. Hasher, L., Goldstein, D., Toppino, T.: Frequency and the conference of referential validity. J. Verbal Learn. Verbal Behav. **16**(1), 107–112 (1977). https://doi.org/10.1016/S0022-5371(77)80012-1
54. High Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI (2019)
55. Hoeyer, K., Hogle, L.F.: Informed consent: the politics of intent and practice in medical research ethics. Ann. Rev. Anthropol. **43**(1), 347–362 (2014). https://doi.org/10.1146/annurev-anthro-102313-030413
56. Hofmann, B.: Suffering: harm to bodies, minds, and persons. In: Handbook of the Philosophy of Medicine, pp. 129–145 (2017)
57. Howard, P., Ganesh, B., Liotsiou, D., Kelly, J., François, C.: The IRA, social media and political polarization in the United States, 2012–2018. U.S, Senate Documents ((2019)
58. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M.: Co-writing with opinionated language models affects users' views. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581196
59. Kang, H., Lou, C.: AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement. J. Comput.-Mediated Commun. **27**(5), zmac014 (2022). https://doi.org/10.1093/jcmc/zmac014
60. Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., Everitt, T.: Discovering agents (2022)
61. Kidd, I.J.K., Medina, J., Pohlhaus Jr., G. (eds.): The Routledge Handbook of Epistemic Injustice. Routledge, London (2017). https://doi.org/10.4324/9781315212043
62. Kligman, M., Culver, C.M.: An analysis of interpersonal manipulation. J. Med. Philos. A Forum Bioeth. Philos. Med. **17**(2), 173–197 (1992). https://doi.org/10.1093/jmp/17.2.173

63. Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. Proc. Natl Acad. Sci. **111**(24), 8788–8790 (2014). https://doi.org/10.1073/pnas.1320040111
64. Krueger, D., Maharaj, T., Leike, J.: Hidden incentives for auto-induced distributional shift (2020)
65. Lavazza, A.: Freedom of thought and mental integrity: The moral requirements for any neural prosthesis. Front. Neurosci. **12**, 82 (2018). https://doi.org/10.3389/fnins.2018.00082.https://www.frontiersin.org/articles/10.3389/fnins.2018.00082
66. Lee, M.K., et al.: WeBuildAI: participatory framework for algorithmic governance. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–35 (2019)
67. Levine, T.R.: Encyclopedia of Deception, vol. 2. Sage Publications, Thousand Oaks (2014)
68. Lewandowsky, S., Van Der Linden, S.: Countering misinformation and fake news through inoculation and Prebunking. Eur. Rev. Soc. Psychol. **32**(2), 348–384 (2021)
69. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2021). https://doi.org/10.3390/e23010018. https://www.mdpi.com/1099-4300/23/1/18
70. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
71. Mahon, J.E.: Contemporary Approaches to the Philosophy of Lying. In: The Oxford Handbook of Lying. Oxford University Press, Oxford (2018). https://doi.org/10.1093/oxfordhb/9780198736578.013.3
72. Martin, C.W.: The Philosophy of Deception. Oxford University Press, Oxford (2009)
73. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: a survey. ACM Comput. Surv. (CSUR) **54**(1), 1–41 (2021)
74. Nissenbaum, H.: Privacy as contextual integrity. Wash. L. Rev. **79**, 119 (2004)
75. Noggle, R.: Manipulative actions: a conceptual and moral analysis. Am. Philoso. Q. **33**(1), 43–55 (1996)
76. Noggle, R.: The ethics of manipulation. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. Summer 2022 edn. (2022)
77. Nozick, R.: Coercion. In: Morgenbesser, M.P.S.S.W. (ed.) Philosophy, Science, and Method: Essays in Honor of Ernest Nagel, pp. 440–72. St Martin's Press, New York (1969)
78. World Health Organization, et al.: Ethics and governance of artificial intelligence for health: WHO guidance (2021)
79. Ovadya, A.: Towards platform democracy: Policymaking beyond corporate CEOs and partisan pressure. https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure
80. Ovadya, A.: 'Generative CI' through collective response systems (2023)
81. Peczenik, A., Karlsson, M.M.: Law, justice and the state: essays on justice and rights. In: Proceedings of the 16th World Congress of the International Association for Philosophy of Law and Social Philosophy (IVR) Reykjavík, 26 May-2 June, 1993, vol. 1. Franz Steiner Verlag (1995)
82. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surv. **55**(3), 51:1–51:44 (2023). https://doi.org/10.1145/3494672
83. Porlezza, C.: Accuracy in journalism (2019). https://doi.org/10.1093/acrefore/9780190228613.013.773

84. Ross, W.D.: Foundations of Ethics. Read Books Ltd., Redditch (2011)
85. Rubel, A., Castro, C., Pham, A.: Autonomy, agency, and responsibility, pp. 21–42. Cambridge University Press (2021). https://doi.org/10.1017/9781108895057.002
86. Rudinow, J.: Manipulation. Ethics **88**(4), 338–347 (1978). https://doi.org/10.1086/292086
87. Sachs, B.: Why coercion is wrong when it's wrong. Australas. J. Philos. **91**(1), 63–82 (2013). https://doi.org/10.1080/00048402.2011.646280
88. Sahbane, I., Ward, F.R., Åslund, C.H.: Experiments with detecting and mitigating AI deception (2023)
89. Sanders, K.: Ethics and Journalism. SAGE Publications, Thousand Oaks (2003). https://books.google.co.uk/books?id=5khuTNSQ6rYC
90. Schlosser, M.: Agency. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab. Winter 2019 edn. Stanford University, Stanford (2019)
91. Schmidt, A.T., Engelen, B.: The ethics of nudging: an overview. Philos. Compass **15**(4), e12658 (2020). https://doi.org/10.1111/phc3.12658
92. Schwartz, M.: Repetition and rated truth value of statements. Am. J. Psychol. **95**(3), 393–407 (1982). https://doi.org/10.2307/1422132
93. E.P. for the Future of Science: Technology: the impact of the general data protection regulation (GDPR) on artificial intelligence (2020)
94. Selinger, E., Whyte, K.: Is there a right way to nudge? The practice and ethics of choice architecture. Soc. Compass **5**(10), 923–935 (2011). https://doi.org/10.1111/j.1751-9020.2011.00413.x
95. Sententia, W.: Neuroethical considerations: cognitive liberty and converging technologies for improving human cognition. Ann. New York Acad. Sci. **1013**(1), 221–228 (2004). https://doi.org/10.1196/annals.1305.014
96. Seymour Fahmy, M.: Love, respect, and interfering with others. Pacific Philos. Q. **92**(2), 174–192 (2011). https://doi.org/10.1111/j.1468-0114.2011.01390.x
97. Shiffrin, S.V.: Speech Matters: On Lying, Morality, and the Law. Princeton University Press, Princeton (2014). https://doi.org/10.1515/9781400852529
98. Spahn, A.: And lead us (not) into persuasion...? Persuasive technology and the ethics of communication. Sci. Eng. Ethics **18**(4), 633–650 (2012). https://doi.org/10.1007/s11948-011-9278-y
99. Srikumar, M., et al.: Advancing ethics review practices in AI research. Nat. Mach. Intell. **4**(12), 1061–1064 (2022). https://doi.org/10.1038/s42256-022-00585-2
100. Sripada, C.S.: What makes a manipulated agent unfree? Philos. Phenomenological Res. **85**(3), 563–593 (2012). https://doi.org/10.1111/j.1933-1592.2011.00527.x
101. Sunstein, C.R.: The ethics of nudging. Yale J. Regul. **32**(2), 413–450 (2015)
102. Sunstein, C.R.: The Ethics of Influence: Government in the Age of Behavioral Science. Cambridge University Press, Cambridge (2016)
103. Taylor, J.S.: Practical Autonomy and Bioethics. Routledge, New York (2009). https://doi.org/10.4324/9780203873991
104. Thomas, S.L., et al.: Young people's awareness of the timing and placement of gambling advertising on traditional and social media platforms: a study of 11–16-year-old's in Australia. Harm Reduction J. **15**(1), 51 (2018). https://doi.org/10.1186/s12954-018-0254-6
105. Thorburn, L., Stray, J., Bengani, P.: Is optimizing for engagement changing us? Understanding recommenders (2022). https://medium.com/understanding-recommenders/is-optimizing-for-engagement-changing-us-9d0ddfb0c65e

106. Tushnet, R.: Chapter 11: Truth and Advertising: The Lanham Act and Commercial Speech Doctrine. Edward Elgar Publishing, Cheltenham, UK (2008). https://doi.org/10.4337/9781848441316.00020
107. UK Department for Science, Innovation and Technology: A Pro-innovation Approach to AI Regulation (2023)
108. US Office of Science and Technology Policy: Blueprint for an AI Bill of Rights (2022)
109. Vold, K., Whittlestone, J.: Privacy, Autonomy, and Personalised Targeting: rethinking how personal data is used. Apollo-University of Cambridge Repository (2019). https://doi.org/10.17863/CAM.43129
110. Véliz, C.: Privacy is Power: Why and How You Should Take Back Control of Your Data. Transworld Digital, London (2020)
111. Wacks, R.: Personal Information: Privacy and the Law. Clarendon Press, Oxford (1989)
112. Waller, M., Rodrigues, O., Cocarascu, O.: Bias mitigation methods for binary classification decision-making systems: survey and recommendations (2023)
113. Ward, F.R., Everitt, T., Belardinelli, F., Toni, F.: Honesty is the best policy: defining and mitigating AI deception. https://causalincentives.com/pdfs/deception-ward-2023.pdf
114. Ward, F.R., Toni, F., Belardinelli, F.: A causal perspective on AI deception in games. In: Proceedings of the 2022 International Conference on Logic Programming Workshops (2022)
115. Ward, F.R., Toni, F., Belardinelli, F.: On agent incentives to manipulate human feedback in multi-agent reward learning scenarios. In: AAMAS, pp. 1759–1761 (2022)
116. Ward, F.R., Toni, F., Belardinelli, F.: Defining deception in structural causal games. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, pp. 2902–2904 (2023)
117. Ward, S.J.A.: Objectivity and bias in journalism (2019). https://doi.org/10.1093/acrefore/9780190228613.013.853
118. Weidinger, L., et al.: Ethical and social risks of harm from language models (2021). https://arxiv.org/abs/2112.04359
119. Wood, A.W.: Coercion, manipulation, exploitation. In: Manipulation: Theory and Practice. Oxford University Press, Oxford (2014). https://doi.org/10.1093/acprof:oso/9780199338207.003.0002
120. Zuboff, S.: The Age of Surveillance Capitalism. Public Affairs, New York (2019)