The Machinery of Power: Artificial Intelligence as a General-Purpose Power Technology

Toby Shevlane and Allan Dafoe, *unpublished*, September 2021 copy.

Abstract

We ask whether artificial intelligence technologies are well-suited to helping actors exercise power over one another, and if so what AI capabilities are most relevant. We define power in terms of one actor having influence over another's behaviour, and we theorise that power can only be exercised through the performance of certain tasks. AI should be considered a general-purpose power technology ("GPPT") because it can perform a range of information processing tasks that are central to the exercise of power across domains. To demonstrate this, we develop a list of interventions that are commonly deployed in the exercise of power: restrictions, incentives, instructions, and persuasion. We show that certain AI capabilities are applicable across all these interventions. Those AI capabilities are: (1) leveraging varied data modalities; (2) identifying individuals; (3) evaluating an individual's behaviour; and (4) predicting the effects of interventions. The general-purpose nature of these capabilities has been underappreciated by social scientists, who often focus on a narrow subset. The lens of AI as a GPPT helps us to understand how AI could lead to structural changes in society.

Introduction

Artificial intelligence (AI) has the potential to fundamentally affect the structure of society. One area where existing digital technologies have already had a profound impact is how actors exercise power over one another. Digital technologies can be found wherever actors influence or control the behaviour of others: the workplace, state regulation, public administration, interpersonal relations, education, law enforcement, and inter-state relations.

This paper assesses the potential for AI to impact how power is exercised. We look specifically at power as defined as the ability of one actor to shape another's behaviour. Is AI especially well-suited to assisting with this kind of power? If so, why — what are the relevant AI capabilities, and how do these go beyond what can be achieved with existing digital technologies?

We argue that AI provides a general-purpose set of tools for assisting with the exercise of power. The economics concept of a "general-purpose technology" refers to a technology that has a pervasive impact across many different sectors of the economy (Bresnahan and Trajtenberg, 1995). By analogy, we argue that AI is a general-purpose *power* technology ("GPPT"). GPPTs significantly alter how actors exercise power across a wide range of contexts.

Other technologies that score highly on this measure would include: writing; the electronic computer; cameras and microphones; and walls, doors, and locks. GPPTs have a deep impact on social life because power relations are key building blocks of social order.

AI has already attracted scholarly and media attention for its bearing on power, especially with respect to surveillance and the manipulation of online behaviour. Nonetheless, AI remains underrated as a GPPT. The stumbling block for many social scientists is that they bring a narrow understanding of AI technologies, failing to identify what is new and interesting about contemporary AI capabilities. Another problem is that scholars consider AI in narrow, isolated domains, such as digital advertising, and this knowledge is rarely aggregated into more general perspectives on AI and power. Our description of AI as a GPPT aims to remedy these failings, drawing attention to a set of widely applicable, power-relevant AI capabilities.

We begin by setting out a theory of power that can incorporate technological change. Exercising power involves the successful performance of a range of tasks — for example, monitoring compliance with standards, or judging what message a recipient will find persuasive. Certain technologies can automate the performance of these tasks. Sometimes this will improve task performance along some important dimension: for example, the power holder can operate with greater scale, cost-effectiveness, precision, rapidity, or context awareness. As a result, the power-holder has a greater potential reach: they can shape a wider range of behaviours, across a larger population, and/or with a higher success rate.

Certain generic tasks play a central role in the exercise of power across different contexts. To understand which technologies will become GPPTs, we must know what these tasks are. We argue that power normally relies upon targeted interventions into an individual's sphere of action. We elaborate four types of targeted intervention, which vary as to the behaviour-shaping mechanism: restrictions (where certain actions are made very difficult or impossible), incentives (where certain actions are rewarded or punished), instructions (where the individual is commanded to take a certain action), and persuasion (where the individual is influenced into voluntarily taking a certain action).

To successfully affect behaviour, these interventions must be responsive to the case at hand, taking in data and making intelligent judgments about what reaction is appropriate. Therefore, three types of task apply universally across interventions: (1) data is collected from the subject; (2) that data is fed into a system of information processing, which decides whether and how the intervention should be applied; and (3) the intervention is implemented (e.g. a message is sent, or a door is locked).

AI becomes a GPPT by performing information processing tasks that have hitherto escaped the reach of computers. We identify several AI capabilities that have universal applicability across different types of power intervention:

- *Leveraging varied data modalities.* AI leverages a much wider range of data modalities than traditional statistical and programming techniques. This is partly because AI performs well on data modalities like images and text that have previously been difficult for machines to meaningfully analyse, and partly because AI allows for easy conversion between different data modalities (e.g. speech-to-text; language translation).

- *Identification*. AI can help with identifying individuals, for example through facial recognition, which is often a necessary part of targeted interventions.

- *Evaluation of behaviour.* AI can be used to evaluate an individual's behaviour against a greatly expanded set of concepts. This is especially useful for judging the individual's behaviour in accordance with a set of rules (as is common under restrictions and incentives)

- *Predicting outcomes*. AI can help with predicting the effect of different variants of an intervention, such that the intervention can be carried out with one eye on its consequences. This is well-known in the case of persuasive interventions, especially online advertisements, where the system is designed to boost click-through rates; but the same principle can be applied across all intervention types.

Within the AI research community, certain capabilities have become associated with power because they have clearly Orwellian associations, such as AI for lip reading or gait recognition. But our analysis suggests that a much greater proportion of AI research is relevant to power. The above capabilities are based upon core research topics: representation learning, regression and classification, reinforcement learning and planning, natural language processing, computer vision, and so on. AI will not become a GPPT thanks to a narrow corner of AI research papers — rather, it is inextricably bound up with the project of building intelligent systems.

The lens of AI as a GPPT helps us to understand an important mechanism through which AI could have a structural impact on society. Nonetheless, in this paper, we do not focus on exploring what those structural changes will be, nor whether they should be welcomed or feared. We also only scratch the surface of the important distributional question of which actors will see their powers increased the most.

Section 1: The task-based view of power

We adopt the classic definition of power given by Dahl (1957, p.202-203): "A has power over B to the extent that he can get B to do something that B would not otherwise do."[1] This definition

[1] NB The emphasis on behaviour-shaping brings Dahl's concept of power very close to how scholars of regulation define the latter. Julia Black, for example, defines regulation as "a process involving the sustained and focused attempt to alter the behaviour of others with the intention of producing a broadly identified outcome" (Black, 2002, p.20) and argues that regulation is practiced by many actors beyond just the state.

emphasises influence or control over somebody's behaviour, i.e. it refers to power *over* rather than power *to*. The definition also excludes "structural power", which is the advantage that certain individuals hold by virtue of their structural position in society, e.g. their gender (Abizadeh, forthcoming) (for structural power and AI, see Benjamin, 2019)**.**

As Dahl (1957) points out, the statement "A has power over B" alone is missing key information: what is the source of A's power, and by what means do they exercise it? Central to our argument is the observation that, in many cases, A must play an active role. For example, A must craft a message that will persuade B; or A must watch over B, judging whether B has breached their orders; or A must reach out and physically block B's movements. In other words, power must be actively exercised through the performance of certain tasks. We refer to this as the *task-based* view of power.[2]

Even under the task-based view, task performance is not everything: it is necessary but not sufficient for the exercise of power. Task performance must be combined with what we refer to as "background power resources", such as the authority to issue legal sanctions, reputation, or money that can be offered as a financial reward.[3]

The task-based view of power helps us to understand how technological innovations might affect power relations. Certain technologies allow for the *automation of power*: some of the tasks necessary for the exercise of power are performed by machines. For example, Latour (1992, 1994) analyses the automation of power (although he does not refer to it as such) in the context of mechanical innovations. He offers various examples: a car that makes a loud noise unless the driver wears a seatbelt; a speed bump; and a door that automatically closes after somebody has gone through. In each case, the relevant artefact performs certain tasks necessary for enforcing behavioural norms which would otherwise require human enforcement.

The same principle applies to information technologies. As an illustration, consider a stamp that automates the process of writing the word "SECRET" on a document. Stamps like this played a role in the control of nuclear information within the Manhattan Project. Wellerstein's (2021) recent history of nuclear secrecy draws attention to the importance of these simple information technologies:

> Even the very mundane aspects of secrecy – like using "SECRET" stamps, required organization. C.P. Baker, a physicist at Cornell, after laboriously hand-marking "SECRET" on every page of a lengthy report in the spring of 1942, left a plea on its final page: "WE NEED A STAMP." (p.40)

[2] The task-based view of power is not especially novel: social scientists (including Dahl, 1957) have long recognised that exercising power involves active measures. But formulating this point in terms of tasks will help us in drawing the connection with AI (and other technologies).

[3] Dahl (1957) makes the same distinction, separating between the "bases" of power, which are "passive", and the "means or instruments" used to exploit the bases.

Leo Szilard, one of the scientists, would later quip that these stamps were “the most dangerous weapon ever invented” (ibid, p.50). More generally, many scholars have drawn attention to the role of paper documents and files, and later the electronic computer, in states’ efforts to regulate populations (Gilliom, 2001; Agar, 2003; Hull, 2012).[4] Foucault (1977, chp.3), for example, emphasises the record-keeping systems that were central to quarantining regimes during outbreaks of the plague in 17th Century France.

We are specifically interested in cases where automation increases A’s power. This is possible because successful automation can improve task performance. Task performance can improve along a range of different dimensions: the task can be performed at lower cost, or with greater scale, speed, accuracy, precision, reliability, or tailoring to the context. Crudely, such improvements could be equated with employing a greater number of humans to work on the task or employing humans of greater skill. The resulting increase in power can take a number of forms: A has power over a greater number of Bs; A has a higher success rate in affecting B’s behaviour; or A has power over a greater range of B’s behaviours.

As an illustration of how automation can increase power, consider the introduction of punched-card machines into early 20th Century state and business bureaucracies (see Agar, 2003, chp. 5). These machines were a precursor to the electronic computer. Information (e.g. a survey response) was physically punched into cards, and machines would systematically sort through those cards. This allowed for more organised storage of data, such as personal records, and for faster statistical analysis of data. For example, in both the UK and the US, the machines allowed new questions to be asked in censuses, both of the population and of business, by increasing the state’s capacity to process the answers. It took the US government seven years to produce tables describing the results of the 1880 census; the punched-card machines were used for the 1890 census, which only took two years, despite being more complicated. For the 1911 census in the UK, the expanded scope for processing census information allowed the state, for the first time, to ask questions about the number of children being born within each household, and to tabulate this against the occupation of the fathers. This was intended to inform future eugenics policies, amid anxiety that the population of the working classes was growing too fast relative to the upper and middle classes.

As Agar (2003, p.152) points out, “The choice of punched-card machinery for the 1911 census was a momentary eugenic spasm, but it was also an anticipation of greater and continuous future data processing by the state.” The punched-card machines were used for a very wide range of bureaucratic tasks: accounting within the emerging welfare state; keeping track of

[4] Hull (2012, 1): “My research began as an exploration of how the Pakistani government shapes social life in Islamabad through its planning and regulatory control of the built environment. However, I gradually came to understand that the modernist program for shaping social order through built forms had expanded a material regime of another, equally significant sort: a regime of paper documents. My conversations with residents about their patches of the built environment of Islamabad quickly veered from family, architecture, and law into stories about the trials and tribulations of their documents and files.”

injuries and disease within the military; making pension calculations for soldiers returning from war; keeping track of local crime information, useful for detective work; and much more. In many domains, the machines automated tasks necessary for producing and organising knowledge about the population, extending the range of activities that could be managed by the state.

Section 2: The landscape of power-relevant tasks

Certain tasks are central to the exercise of power. GPPTs exist not only because there are technologies that can perform many different tasks, but (more importantly) because there are certain generic tasks that are involved in the exercise of power across many contexts. To understand what technologies will become GPPTs, we must be able to identify these generic, power-relevant tasks.

We look at a range of interventions that A can use to affect B's behaviour: A can restrict, incentivise, instruct, or persuade B.[5] Each of these interventions is described in detail below. This list is not exhaustive, and we briefly note a few additional interventions at the end of the section. We would argue that these kinds of interventions are necessary for the exercise of power, at least under the definition of power given above. A needs some mechanism for having a deliberate impact on B's behaviour, and these interventions all constitute standard means of doing so. They are all widely deployed, often in combination, across various domains: law enforcement, the regulation of firms, the management of employees, and so on. We refer to them as *targeted interventions.*

Despite their differences, targeted interventions all rely on certain kinds of task. First, for the intervention to be effective, it must be grounded in the empirical reality of the situation (e.g. what is B doing?). Hence, there is always a need for *data collection*. Second, *information processing* is required, both in order to make sense of the collected data, and to determine the appropriate response. Third, the intervention must be *implemented*.[6] For example, in the case of a persuasive intervention, A sends B a message; or in the case of an incentive-based intervention, A might send B a financial reward. This third group of tasks is more open-ended than the other two, mapping less neatly onto specific technologies (although sending information over long distances is often required; and the application of physical force is also sometimes relevant).[7]

Existing GPPTs can be explained by reference to this framework. For example, the camera earns its status as a GPPT through data collection. The punched-card machine became a GPPT through information processing (including information storage, and statistical analysis)

[5] Bertrand Russell (1938) offers a similar typology.
[6] This tripartite framework is similar to that of Fourcade and Healy (2017), who - in the context of data-driven markets - separate between the "dragnet" that collects data, the "scoring" process (e.g. assigning creditworthiness), and the intervention in the behaviour of the user.
[7] Because we are especially focussed on AI technologies, we do not break down this category of tasks in much detail.

and perhaps to a lesser extent through data collection (insofar as punching a card is easier than ticking a box). The electronic computer is very dominant as a mode of information processing, but also helps with data collection (e.g. by collecting user data) and the implementation of interventions too (e.g. as a medium for sending messages). The Internet is also relevant to all three types of task.[8] We claim that AI is a GPPT specifically due to its potential in the area of information processing.[9]

In what follows, we describe each intervention — restrictions, incentives, instructions, and persuasion — and identify the relevant information processing tasks that AI could perform. Later, in section 3, we will bring together and systematise the insights about AI capabilities.

A. Restrictions

A restriction-based intervention occurs when an individual's actions are selectively restrained, based on a judgment about the individual, their actions, the possible effects of those actions, or any other situational factor (see Kerr, 2010; Brownsword, 2015). Restrictions do not rely on incentives or persuasion; instead, certain actions are simply made impossible (hard restriction) or very difficult (soft restriction).

These restrictions normally rely on a mix of human and machine labour. This has been true for many years, although technological progress has led to greater automation. At the city gates of medieval European cities, human gatekeepers would ask standard lists of questions to arriving travellers, turning their recent history into data that could be processed in accordance with the rulebook about who was authorised to enter the city (Jutte, 2014). The city walls allowed this interrogation to take place, by forcing arrivals to enter through the gates, and the gates themselves allowed for selective entry. Although in medieval times, human gatekeepers were needed to judge who was allowed to enter, today, the bundle of tasks carried out by the gatekeeper is often partially automated. At the modern airport, a machine scans the contents of travellers' bags, and machines, as well as humans, make sense of that information.

In some cases, restrictions are imposed by an artefact on its user. In December 2016, a terrorist attack took place in Berlin, where a man drove a hijacked truck into a Christmas market, killing 12 people. The truck's movements were erratic, and it came to a stop unexpectedly early. This prompted early speculation that the terrorist was wrestling with the truck driver during the attack. However, it was later revealed that he was wrestling with something else: the truck's automatic

[8] Data is collected over the internet; information processing can be distributed across multiple, networked locations; and interventions can be communicated over the internet (e.g. an employer's instructions can be communicated via email or video call).

[9] In the context of AI, we need to clarify the role of data by drawing a distinction between the use of AI systems and their training. In the above schema, "data collection" would refer to situation-specific data (such as data about the target individual) which is then inputted into an AI system for analysis. This is different from training data, which should be considered a background resource, often useful for improving the capabilities of an AI system. This training data might closely match the intended use case, e.g. where click data from social media users is used to train recommender systems that then recommend content back to a similar pool of users. But this is not necessarily so: for example, a text classifier could be pre-trained using publicly available, generic datasets, such as those scraped from Wikipedia, and then fine-tuned on a hand-labelled dataset of statements.

braking system, which had perceived the imminent collisions (Taylor, 2016; Dávideková and Greguš, 2017).[10] Such systems narrow the range of ways in which cars and trucks can be used. As with this example, the restriction often involves distinguishing between different types of action. In the same way, Facebook Messenger blocks the sending of certain prohibited URL links;[11] locationized guns will only shoot in particular geographical zones; and there are new models of defibrillator that will only administer a shock when necessary, based on the pattern of the recipient's heartbeat. In other cases, the restriction involves identity-based distinction, targeting individuals with certain characteristics. Televisions, like screw-capped bottles, can be "child-locked"; many mobile phones are fingerprint locked.[12]

Computerised devices are well-equipped for stepping in and imposing restrictions. For example, with Facebook Messenger, the sending and receiving of the message is handled by software, which has ample opportunity to filter the content of the message (e.g. censoring certain URL links). In contrast, the telegraph has no mechanism for selectively blocking certain types of message: the electricity simply travels down the wires, as electrical pulses to be interpreted by the human on the other end. Another example: in 2009, Amazon realized that they had mistakenly sold ebook copies of George Orwell's *1984* and *Animal Farm*, and simply deleted the books from users' devices (Kerr 2010). In contrast, sellers of hardcopies cannot recall books at such scale and convenience.

The ability to impose restrictions must be combined with a system for deciding when and how a restriction should be imposed. Existing computer software has already made progress on this front, but such software has been limited in the kinds of analyses that can be performed automatically. Does the entered password match the password on file? Is this smartphone the primary device for this music file? Is the geo-location of this gun outside of the permitted coordinates? Or in the case of the defibrillator: what is the variance of the time intervals between heart beats, and when that is combined with other, similar measures, does the ECG reading cross the threshold of shock-worthiness? These analyses all draw upon competences that computers have had for decades now, such as: storing information in databases and searching for matches on those databases; keeping track of time; making arithmetic calculations; and plugging numerical values into regression models.

[10] The truck was a Scania R 450. See here for a November 2013 description of Scania's automatic braking system. Note that it would be possible for the driver to disable or override the system.
[11] For a discussion of "algorithmic censorship", see Cobbe (2020).
[12] Most of these examples fit the description, given above, of data being collected (e.g. on how an artefact is being used) and then fed into an information processing system. However, an exception is the child-lock on a screw-top bottle. Here, it might seem like a stretch to imagine the bottle "collecting data" on how the lid is being twisted, and computing some judgment about the user's manual dexterity or knowledge. Rather, the designer has found a more mechanical method for drawing distinctions between individuals: a mechanical test, where failure correlates strongly with being a young child. However, often the distinction between permitted and non-permitted actions cannot be erected mechanically, and so more sophisticated information processing will be required.

AI expands this toolkit. The basic structure of the workflow stays the same: data is inputted, then that data is processed and analysed, culminating in some algorithm for resolving the binary question of whether the restriction should be applied. At a very general level, AI expands the range of classifications and regressions to which the data can be subjected. A classification task involves imposing categorical classes on the data (e.g. categorising text as “hate speech” or not), whereas a regression task outputs a quantitative score instead (e.g. assigning a 0-100 score for the offensiveness of a comment on an online forum). Deep learning systems can perform such analyses both: (a) across a wider range of inputs, making sense of important data modalities such as images, videos, text, and audio, and (b) by reference to an expanded range of concepts, i.e. the data can be projected onto a more semantically rich space.

As such, deep learning provides a general-purpose set of tools for analysing the content of an individual’s behaviour. For example, instead of picking out key-terms, modern natural language processing techniques can perform more qualitative analyses: the politeness of an email (Madaan et al, 2020); whether a tweet is a political parody (Maronikolakis et al, 2020); or whether a social media post appears suicidal (Shing, Resnik, and Oard, 2020). Video recognition systems can decipher whether individuals are keeping a two-meter distance from each other during an epidemic (LandingAI, 2020); or they can monitor for abnormal activity in multi-storey residence buildings (Jia et al, 2020). Such assessments may suffer from problems of validity, accuracy, and bias. Nevertheless, they will often be sufficiently functional to serve the interests of those designing the targeted intervention. In the case of restrictions, deep learning expands the range of automated analyses that can condition whether or not the restriction is imposed.

B. Incentives

Whereas a restriction makes certain actions more difficult, an incentives-based intervention makes actions more or less appealing: costs and benefits are conditioned on the performance of certain actions.[13] These incentives can take any form, such as money, legal punishments, or shame. They operate prospectively, in that the individual modifies their behaviour in expectation of incentives that will be applied in the future.

The ubiquity of computerised information technologies has facilitated novel incentive-based interventions. Uber drives are monitored for the percentage of journey requests that they accept; a driver whose trip acceptance rate falls below a certain percentage can be automatically suspended (Rosenblat, 2018, p.150). During the Covid-19 pandemic, many countries have deployed smartphone-based apps for policing compliance with stay-at-home orders. These apps monitor the individual’s GPS location, or request that the user takes a photo of their environment. Non-compliance can lead to punishment, such as a fine. In education, a popular software tool used to monitor children’s classroom performance uses a points-based

[13] Technically, there is an overlap between these categories: making an action very difficult (qua a targeted restriction) will normally simultaneously increase the costs of performing that action. For our purposes, these cases can be excluded from the category of incentives.

rating system, alongside various qualitative categories such as “displaying grit” (Manolev et al, 2019, p.40).

The necessary tasks can be broken down into the same three categories as above: data collection, information processing, and the implementation (in this case, the “intervention” is where the incentives are dished out). For example, with a speed camera, the machine bounces radio waves off oncoming vehicles (a method of data collection). Then, as information processing: the internal computer analyses those radio waves to compute distances, which are converted into speeds. A decision rule is applied whereby speeds over a certain limit will be sanctioned. Further data collection is thereby triggered: an image is taken of the vehicle. Then comes another string of information processing tasks: computer vision techniques are used to decipher the car’s licence plate (and note that the plate was itself performing the task of broadcasting such information). These numbers and letters are then sent over the internet to a centralised server, and entered as a search term across a database of drivers. Finally, to implement the sanction, a letter communicating the fine is generated and sent to the driver’s home address via the postal service.[14] It is worth noting the wide range of information processing tasks necessary to make speed cameras work, which includes: arithmetic calculations, computer vision, sending information between networked computer systems, search functions, and database management.

The information processing tasks aim to establish what happened (who, what, when, where, why?) and then to convert these factual conclusions into a decision about what incentives should be applied. This is analogous to how a court must deal both with questions of fact (what happened?) and questions of law (how do the rules apply to these facts?). For the speed camera, most of the work is in establishing the facts; subsequently mapping those facts onto the rulebook is then relatively simple — e.g. in pseudocode: if (speed > 35) then Print “£100 FINE”, else Print “NO SANCTION”. In other domains, such as the courtroom, applying the rulebook is itself a tricky exercise. In the example of judicial decision-making, the difficulty comes not only from the fact that legal rules are often complex, but also because they involve loosely specified categories such as “reasonable care”, the application of which requires background knowledge and strong reasoning abilities. The same is true for the application of social norms, which rarely involves following computer-friendly, numerical procedures like the application of speed limits.

The underlying information processing tasks are highly overlapping with those for restrictions, above. This is despite the fact that incentives differ from restrictions in various ways: they are retrospective; the decision-space is more expressive, in that incentives can be graded (e.g. larger fines for higher speeds) and multi-dimensional (e.g. the fine is accompanied by “points” deducted from the driver’s licence); and there are differences in the portfolio of background resources that will enable the power-holder to apply incentives (for example, financial wealth is

[14] NB the ability to successfully perform this intervention is reliant on the legal authority of the local authority to issue such fines. In this example, law is not replaced by technology - rather, both are foundational to the targeted intervention.

required for handing out financial rewards, and a certain level of authority is useful for meting out reputational penalties).

Nonetheless, there are certain generic information-processing tasks that can be repeated. Incentive-based interventions are highly amenable to automation by the kinds of AI advances that assist in evaluating human behaviour: the expanding range of classifications and regressions that AI systems can be trained to perform, and the expanded range of data modalities that can be leveraged. Computers are no longer restricted to processing the easily quantified aspects of human behaviour — such as an employee's customer satisfaction ratings or the number of hours they work — but can make more qualitative judgments too. Is this person part of the rioting, or just trying to visit nearby shops? How polite is this employee when interacting with clients? As such, the use of AI in shaping behaviour should not be conflated with "governance by numbers" (Supiot, 2015) and the sociology of quantification more generally. Neither the data fed into the AI system, nor the concepts it imposes, must appear quantitative in nature; and the intervention need not involve quantitative scoring or ranking of individuals — although these are all possible. This is a key way in which AI now provides a more flexible, general-purpose technology for executing targeted interventions than traditional computing technologies.

In addition, AI provides new methods for identifying individuals, e.g. through analysis of faces, voices, and walking styles. The problem of identifying individuals is especially salient for incentives, because knowing the individual's identity is often a necessary step in rewarding or punishing them. This is in addition to how (as with restrictions) identification can be used for applying standards of behaviour that make different demands of different individuals — for example, quarantine rules that apply to individuals who have recently tested positive for a virus.

C. Instructions

The third category of intervention occurs where A gives B instructions about what action to take, based on real-time decision-making, taking into account the contingencies of the situation at hand. Under this category of interventions, we assume that B will comply with A's instructions — or at least, we treat that question as exogenous. Compliance could be secured, for example, by systems of incentives (see above), or the authority of the instruction-giver (be that cultural, bureaucratic, or personal authority: Weber, 1921). Therefore, the relationship of employment is the archetypal setting for this type of intervention.

Modern examples of instructional interventions include the Uber app, which allocates drivers to particular fares depending on the driver's position. Project management software also facilitates instruction-giving. The software displays to the manager the calendars of the employees, and how utilized their time is across different days, and from this view the manager can allocate the employees to particular projects. The software makes the employee's activities more "legible" (Scott, 1998; Foucault, 1977), allowing the manager to see where to intervene.

If we are already assuming that B will follow A's instructions, then how do technological improvements lead to an increase in A's power? The starting point is that automation can make A's role as an instruction giver easier or increase the quality and relevance of A's instructions. Therefore, if A is an existing employer of B (or has some similar relationship), it might become in A's interests to take a more hands-on role in directing B. We can see this process in the history of the telegraph (see Nickles, 2003, chp.2). Before the telegraph, diplomats negotiating abroad could not easily check back in with their home governments, and so would need to be granted high levels of autonomy to make international agreements.[15] The telegraph meant that the home government could be kept in the loop, and governments thereby established more fine-grained control over negotiations.[16] A similar change took place with naval captains. Nickles (2003, p.43) quotes a US admiral: "The cable spoiled the old Asiatic Station. Before it was laid, one really was somebody out there, but afterwards one simply became a damned errand boy at the end of a telegraph wire." In these cases, the telegraph led to an increase in the power of central government, both because it became easier for the central government to give instructions, and because those instructions could be higher quality (because the government was better informed about the situation on the ground).

Even if A is not already B's employer, if A has a newfound ability to direct B's behaviour cheaply and fruitfully, then there might be structural pressures for A to become B's employer (or something like it). This seems to explain Uber and other similar applications, which have moved from non-existence to directing, at a fine level of detail, the activities of a very large, international fleet of cab drivers.

Again, instructional interventions rely on information processing: A must obtain information about B's situation, and intelligently process that information to make judgments about what instructions to give. This means that, as well as communication technologies like the telegraph, information technologies can assist with these interventions. A range of cognitive tasks must be performed: perception of the situation facing B (e.g. making sense of a video feed), predicting future developments, evaluating different strategies, planning, and communicating effectively. These tasks are amenable to automation through continued progress in AI technologies. Relevant AI topics include the broad category of "perception" (e.g. image recognition, or the conversion of speech to text), and the broad camp of techniques relevant to automated, real-time decision-making, such as reinforcement learning. Improvements in these areas lower the costs of power-holders scaling-up their instruction-giving operations, and increase the range of situations in which decision-making will be more efficiently carried out by a central hub.

[15] Nickles (2003) gives the example of the USA's purchase of New Orleans from France. President Jefferson said that no set of instructions could be "squared to fit" the contingencies of the negotiation.
[16] A 1900 New York Times article, extracted by Nickles (2003, p.45), argued that the diplomat: "has become less of a statesman and more of a correspondent, an exponent of his master's views, a go-between, an instrument."

D. Persuasion

Persuasion-based interventions occur where A composes a message and sends it to B, and the message has some persuasive effect on B. This kind of ability to shape what people know and think has long been considered a form of power. As Lukes (1974, p.23) put it: "...*A* may exercise power over *B* by getting him to do what he does not want to do, but he also exercises power over him by influencing, shaping or determining his very wants."

Persuasion can be carried out person-to-person, through speech, but in practice, these interventions are often reliant on various technologies. Such reliance on technology has long been the case: consider the Medieval kings whose messengers would ride on horseback to spread the message faster. Modern technologies automate the process to a greater extent. For example, in the case of a modern advertisement: the content of an advertisement, and its audience, may be tailored to specific individuals' needs and wants, which have been captured during their internet browsing; the content of the message can be updated in response to feedback signals coming from the potential customers; and the communication of the message takes place through computer systems connected via the internet.

Modern technologies have changed the way that states communicate with the population. For example, states send emails reminding individuals to pay their taxes, tailoring the message to the individual, and sometimes updating the message based on feedback (e.g. Behavioural Insights Team, 2018). In March 2020, during the Covid-19 pandemic, the UK's National Health Service reported that:

> daily text messages are being sent to over 1 million people who have been identified by the NHS in England as needing to protect themselves by self-isolating for at least 12 weeks because they are extremely vulnerable to COVID-19. This group includes people who have had organ transplants, have certain types of cancers, or have significant respiratory conditions. (Smith, 2020)

As this example demonstrates, the management of the health of the population involves targeted outreach, relying on modern communication technologies, and knowledge of individual characteristics, stored on computerised databases.

Political propaganda is also being automated and extended by modern information technologies. Political scientists have begun to study "computational propaganda" (Woolley and Howard, 2018), which is the "use of algorithms, automation, and human curation to purposefully manage and distribute misleading information over social media networks" (p.3). This involves the use of social media "bots", which are computer programmes designed to engage in online debates. In the case of computational propaganda, as it is defined by Woolley and Howard (2018), the posts are misleading, but the more general phenomenon of machine-assisted, large-scale persuasion need not be confined in this way.

Artificial intelligence is likely to allow targeted messaging to be automated to a greater extent, potentially increasing both its scale and effectiveness. At the present cutting edge of AI research, AI systems are able to generate fabricated news articles that humans cannot identify as machine-written (Brown et al, 2020). The ability to direct the outputs of such text-generating AI systems, e.g. toward arguing for a particular position, is currently a research direction within the field (Keskar et al, 2019; Ammanabrolu et al, 2020). Aside from automating the composition of messages, AI is also relevant to the task of modelling who should be targeted and what messages different recipients will find persuasive — for which digital advertising is the archetype. States and companies already tailor messages to specific groups, and attempt to increase the level of persuasion through greater knowledge of the recipient. Advances in AI will further this aim, by allowing more to be known about the recipient and how they will respond to a given message.

E. Additional types of targeted intervention

We have not covered the full space of targeted interventions. Additional examples of targeted interventions are:

1. *Recommender systems*. These are systems that determine the content to be displayed to a user, such as news articles, answers to queries, or social media posts (Milano et al, 2021). They can be designed so as to achieve certain effects on user behaviour, such as to boost their engagement, or to have a persuasive effect. Recommendations differ from persuasive messages (above) in that the content is not produced by the recommender.
2. *Nudges*. Nudges are interventions that make certain choices slightly easier or more difficult (Thaler and Sunstein, 2008). A nudge is thus similar to a restriction, but weaker. Nudges are widely used in the digital world as a way of influencing behaviour (Yeung, 2017).
3. *Selective disclosure*. This is where one actor decides whether to reveal or withhold some information from another, knowing that each option will have a different effect on the recipient's behaviour. An example is how participants in drug trials are not told whether they have been given the treatment or placebo. This is designed to ensure that the behaviour of those in the treatment and placebo group is functionally the same, when otherwise it would not be.

These all adhere to the basic framework of tasks, requiring data collection, information processing, and then the relevant intervention. Again, they also require certain background resources: for example, selective disclosure requires possession of private information.

Section 3: What is special about AI? AI as a general-purpose power technology

AI provides a general-purpose set of tools for automating an important set of power-relevant tasks. This is partly because information processing tasks are so central to exercising power, combined with the fact that AI has the potential to automate so many different types of information processing task. In this sense, AI is not one technology, but many. We have already introduced a number of power-relevant AI capabilities, such as the ability to identify individuals and the ability to classify behaviour. This section brings together, and further develops, this set of AI capabilities. This is an important exercise because social scientists studying AI often focus on a narrow slice of these capabilities, and thus fail to fully appreciate the significance of AI for power.

Four power-relevant AI capabilities can be identified, as follows. Each is generally relevant across multiple types of targeted intervention. This list could be longer (e.g. it does not include robotics, or the generation of text and images), but we focus on the capabilities that have the most general applicability.

(1) *Leveraging varied data modalities.* Compared to traditional computer systems, AI systems can now analyse, and convert between, an expanded set of data modalities. This capability is very significant for targeted interventions, which rely on data being fed into an information processing system and are therefore bottlenecked by the types of data that can be processed by such systems.

The expansion in data modalities partly comes from successes in fields working on specific modalities, such as natural language processing (NLP) and computer vision. But importantly, certain generic techniques have proven successful in modelling a vast range of different data modalities. The clearest example is the Transformer model (Vaswani et al, 2017). This architecture first came to prominence in NLP, with models like BERT and GPT-2, which are pre-trained on large corpuses of unstructured text data scraped from the web (Devlin et al, 2018; Radford et al, 2019). These models can then be fine-tuned on specific tasks, such as text classification. The important point is that the same architecture and training approach has subsequently proven successful in modelling other modalities, including images (Chen et al, 2020) and DNA (Ji et al, 2020). Chen et al (2020) argue that “Transformer models like BERT and GPT-2 are domain agnostic, meaning that they can be directly applied to 1-D sequences of any form.”

In addition, AI is useful for converting between data modalities. One key example is speech-to-text methods, which convert audio data into text. That text data can then be analysed using NLP techniques. Lip-reading techniques are similar in that, where possible, they convert video data into text. Another example is machine translation, which converts between different languages. This is useful if the human or AI system analysing that data can only work with a

specific language (and hence translation work has historically been valued by colonial rulers: Cohn, 1996).

(2) *Identifying individuals*. AI techniques can be used to help narrow down somebody's identity, based on their voice, their face, the writing style, their walking style, and so on. Targeted interventions must often be targeted towards specific individuals — for example, because: (a) a set of rules discriminates between different groups of people, such as restrictions on how long children can play video games; (b) as we saw with targeted incentives, somebody specific must be given the rewards or punishments; (c) an instruction must be given to a specific subordinate, and a persuasive message must be sent to a specific group of individuals (e.g. swing voters). Of the four AI capabilities listed here, identification has already attracted a certain amount of attention for its relevance to power, and as such is overrated relative to the other three. (Identification also has less room to grow as AI capabilities increase.)

(3) Evaluation of behaviour through classification and regression. This is the ability to impose concepts on behavioural data, either by sorting the behaviour into categories or by scoring it along some dimension. As we have already argued, this process is fundamental to targeted restrictions and incentives, which both often rely on judging behaviour against certain criteria. Supervised learning allows the designer of the targeted intervention to specify, by reference to labelled examples, which behaviours they wish to restrict, reward, or punish - although current methods require a lot of training data. As we argued above, these methods provide a platform for teaching machines a much broader range of concepts than was previously possible. This goes far beyond the sorting of individuals on the basis of personal characteristics such as income and gender, as has been done for many years, e.g. within consumer credit scoring.

Furthermore, the evaluation of behaviour is also relevant to both instruction-based and persuasive interventions. Both these interventions often rely on A forming an understanding of B's situation: a supervisor benefits from knowing what their employee is doing and how they are performing; and the information processing required for a targeted advertisement, for example, could include compressing an individual's online behaviour into meaningful representations.

(4) *Predicting how individuals will respond to different interventions*. This is where the AI system is used to select what form the intervention should take, by predicting how the individual will respond to different variants. This capability is often (narrowly) associated with targeted advertising on social media — and therefore persuasive interventions. Here, the individual's personal profile is the independent variable, including data on their online behaviour, and the role of deep learning is in statistical analysis of how, in large datasets, these personal datapoints are associated with measures of engagement such as clicking on an advert. This paradigmatic case carries certain features, none of which are essential: (a) the data used to train the model takes the same form as the individual's data subsequently fed into the model for predictions; (b) the data is obviously personal in nature; (c) the data is monopolised by certain large firms; (d) the aim of the intervention is to persuade or boost engagement.

The fundamental principle — of using AI to predict the effect of different forms of intervention — can be generalised to other categories of targeted interventions, including restrictions and incentives. So far, we have primarily considered the latter interventions in a rules-based context, where behaviour is evaluated against a fixed standard. The speed camera, for example, is delegated the task of enforcing a particular speed limit, but is not responsible for considering the link between speed limit enforcement and the outcome that really matters, which is the number of accidents in an area. This responsibility still resides with the local government, which must inform itself as to whether speed cameras are effective at reducing accidents, and must set the parameters of the intervention — where the camera is placed, and what speeds should trigger sanction — in accordance with the policy goal of reducing accidents (and perhaps other policy goals, too).

However, in future, we could imagine more agentic systems, which flexibly select the form of the intervention in line with some goal (Yeung, 2018). A software *agent*, such as a chess programme, is one "situated within and part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to affect what it senses in the future" (Franklin and Graesser, 1997; see also Brustoloni, 1991). Software is more agential to the extent that it: (1) flexibly selects from a wide range of possible moves; (b) uses planning techniques; for example, a chess programme looks ahead and considers chains of moves and responses; (c) updates its policies over time in light of what works. The example of automated braking is a step in this direction. The system does not enforce a fixed standard about how best to drive safely, but rather applies the brakes with the timing and force required to avert foreseen collisions.

Finally, we have so far considered AI systems that consult very narrow, context-specific models of the individual's behaviour (as with targeted adverts) or their local environment (as with automatic braking). Over the long term, it is possible to imagine AI systems that consult a much richer, more general model of humans and their environment. Some people have argued that very large language models like GPT-3 are beginning to develop sophisticated "world models", i.e. a general understanding of the way the world works, although there is currently much room for improvement (Gao, 2020). If this proves possible, such models could play a role in predicting the outcomes of targeted interventions. Moreover, there are signs that these very large, general models are more data efficient when retrained on data from specific contexts (Kaplan et al, 2020), which could enable a greater range of actors, beyond those with large, private datasets, to target interventions on the basis of predicted outcomes (see Tucker et al, 2020). Overall, then, the potential significance of this category of AI capabilities goes well beyond the paradigmatic case of social media advertising.

[End list]

We are now in a position to describe the intersection between (a) the set of tasks that AI performs, and (b) the set of tasks that are necessary for the exercise of power (see Figure 2). As a general-purpose technology, AI has a broad range of potential applications, many of which

are not closely tied to the exercise of power. However, AI has a special connection with power. Exercising power requires the performance of a specific portfolio of tasks, with information processing playing a very central role. Within information processing, AI covers not just one, but several tasks that are universally applicable across a wide range of instances in which power is exercised. This is what motivates AI's status as a general-purpose technology for exercising power.

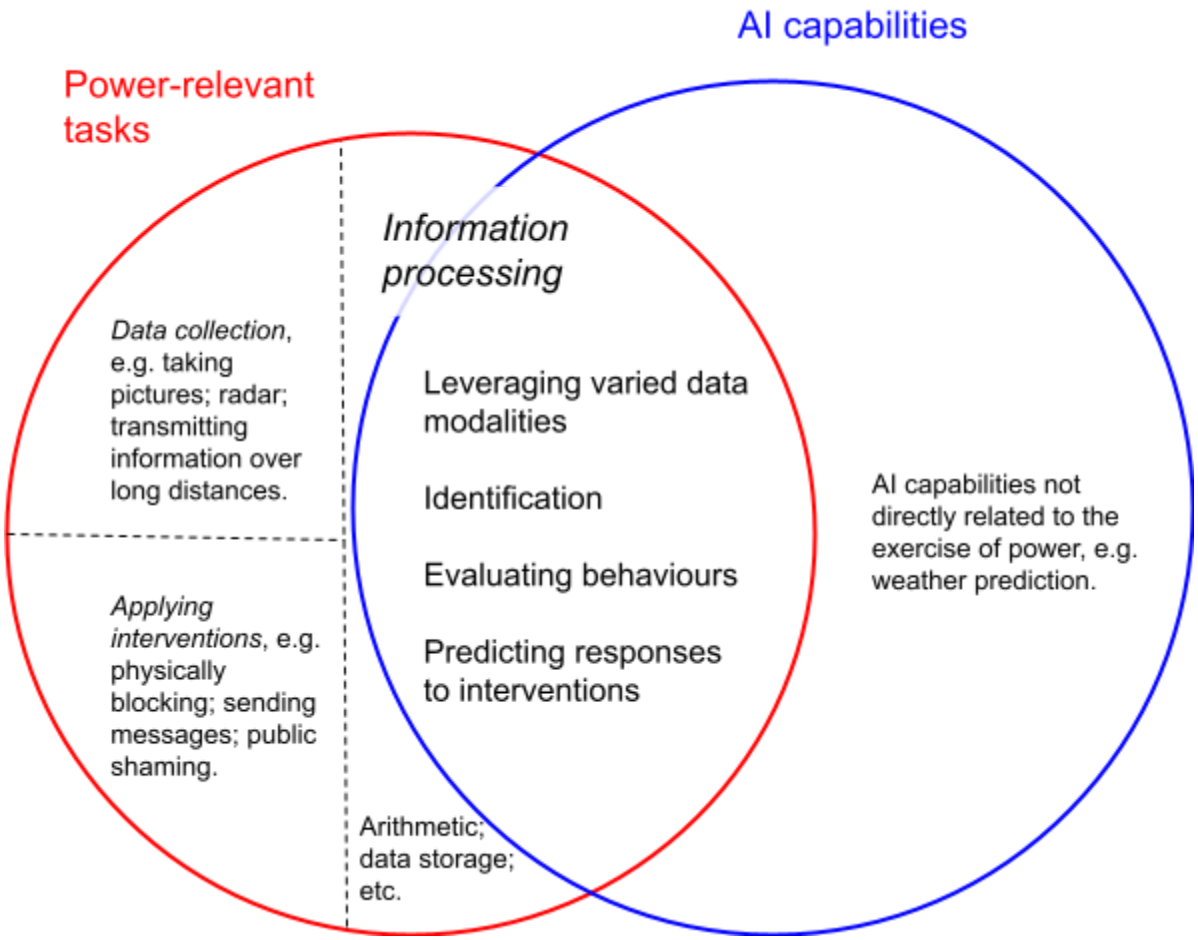
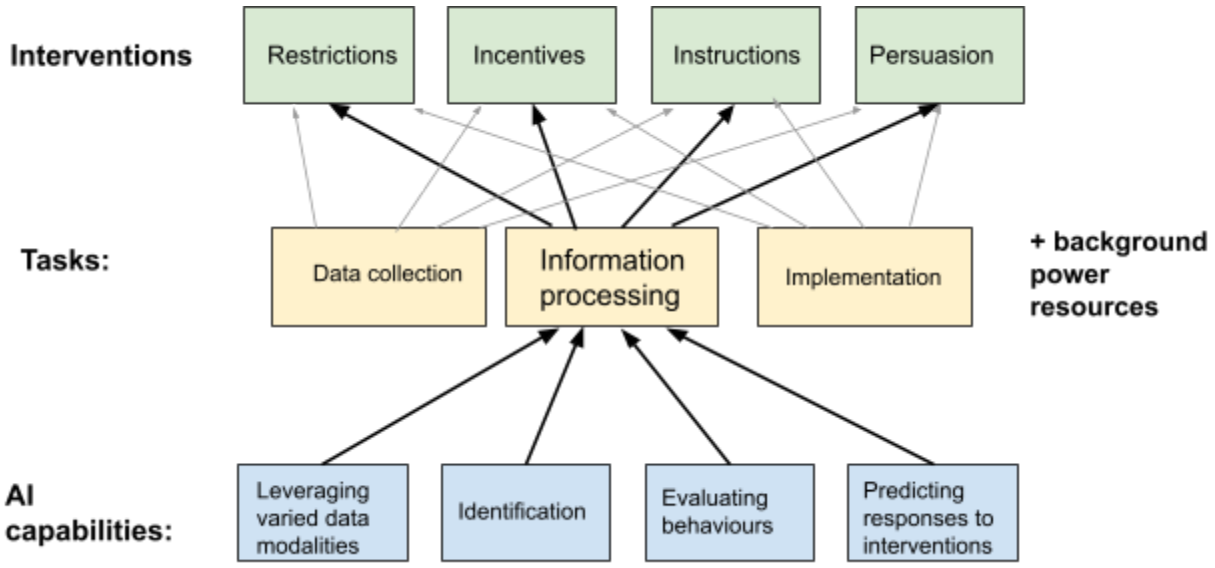


Figure 2: The intersection between AI and power-relevant tasks.

Alternative diagram:



To reach this conclusion, we have drawn upon a wide sample of AI research topics. There is a certain fallacy within the AI research community that says: the landscape of AI research papers contains a small corner labelled “surveillance applications”, which covers papers on, for example, lip-reading and facial recognition for oppressed populations. However, AI does not become a general-purpose power technology thanks to these small, remote outposts — rather, the central ambitions of AI research are directly relevant to power.

Conclusion

We have demonstrated why AI should be considered a general-purpose power technology. AI has the potential to automate several important categories of information processing tasks that are central to the exercise of power. Although AI is a continuation of existing digital technologies, AI pushes out the boundaries of what’s possible in important ways, and will therefore have its own effects on power.

One criticism might be that, in seeking to articulate a theory of power that accommodates AI technologies, we have distorted how power actually works. Our focus on targeted interventions emphasises a kind of power that is relational, intentional, and grounded in micro-level activity. This view neglects, for example, the role of ideologies or “systems of thought” (as found in the work of Marx, Bourdieu, and Foucault’s earlier works). In the same way, some readers might argue that our exclusion of “structural power” (Abidezeh, forthcoming) was too costly. In defence, we would maintain that the vision of power that we rely upon here, even if not fully comprehensive, still represents a very widespread and important social phenomenon.

A related issue is whether our focus on tasks obscures the role of power resources like money, legal authority, social status, and the “platform power” that technology companies hold by virtue of many users accessing their services. We would argue that the task-based view of power

complements, rather than competes with, a focus on these other power resources. Task performance is necessary but rarely sufficient for the exercise of power. Certain actors, such as states and large firms, are especially well-placed to carry out targeted interventions. These actors are best placed to collect data on individual behaviour (Pasquale, 2015), and have an outsized share of the background resources needed to apply interventions, such as the legal authority to punish individuals. The diffusion of AI capabilities, therefore, will not necessarily diffuse power, and in many areas will act as a force multiplier for existing forms of authority.

Another potential criticism is that we have presented a rose-tinted view of what AI technology can achieve. Scholars studying the impact of AI often focus on its failings: AI systems can be biased, and there are many decision-making contexts, such as recruitment, where AI systems cannot match the validity of human judgment. We would not dispute this, but we would maintain that a full understanding of the impact of AI requires paying attention to what happens *both* when AI works and when it does not. (By analogy, the societal impact of automobiles flows through not only road accidents and CO2 emissions, but also the fact that people can travel further, faster, and more conveniently.) We also want our analysis to be robust in the face of continued progress in AI capabilities. The AI capabilities that we identify have not yet been maximally realised by the current state of the art — AI still has room to grow as a GPPT.

Going forward, one of the most important governance challenges of our age is to shape how technology-enabled power is designed and implemented. Insufficient attention is currently directed toward this problem. Langdon Winner's critique from 1986 still bites: we are technological somnambulists, sleepwalking "through the process of reconstituting the conditions of human existence" (Winner, 1986, p.10). This description applies to the continued progress in AI capabilities, which — by reshaping relationships of power in society — will alter the fundamental building blocks of social order.[17] The lens of AI as a GPPT sheds light on the deep and pervasive impact that AI could have, which is a precondition for well-informed governance of the technology.

[17] In particular, we would highlight that technology-enabled power has an important and complicated relationship with existential risk to humanity's future. Ord (2020, p.154) highlights the risk of "unrecoverable, enforced dystopia", where thanks to technologies like AI, a very stable totalitarian regime achieves "global dominance and absolute control, locking the world into a miserable condition." On the other hand, Bostrom (2019) argues that an increase in technology-enabled surveillance and social control would be necessary to protect against some existential risks, such as certain engineered pandemics. As these examples demonstrate, advances in AI will supercharge a policy tension that has already been front-and-centre in recent decades, which is how the governance of digital technologies should balance liberty and security.

References

Agar, J. (2003). *The government machine: A revolutionary history of the computer*. MIT Press.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction*. Oxford University Press.

Ammanabrolu, P., Urbanek, J., Li, M., Szlam, A., Rocktäschel, T., & Weston, J. (2020). How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds. *ArXiv:2010.00685 [Cs]*. http://arxiv.org/abs/2010.00685

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 759–760. https://doi.org/10.1145/3041021.3054223

Beniger, J. R. (1986). *The control revolution: Technological and economic origins of the information society*. Harvard University Press.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Harvard University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs.CL]*.

Brownsword, R. (2015). In the year 2061: From law to technological management. *Law, Innovation and Technology*, *7*(1), 1–51. https://doi.org/10.1080/17579961.2015.1052642

Calo, R., & Citron, D. K. (2020). The Automated Administrative State: A Crisis of Legitimacy. *Emory Law Journal, Forthcoming*.

Chandler, A. D. (1977). *The visible hand: The managerial revolution in American business*. Harvard University Press.

Cobbe, J. (2020). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00429-0

Dahl, R. A. (1957). The Concept of Power. *Behavioral Science*, *2*(3), 201. Periodicals Archive Online; Periodicals Index Online.

Dávideková, M., & Greguš, M. (2017). Nice, Berlin, London—If every car had autonomous emergency braking system for forward collisions avoidance. *14th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2017) / 12th*

International Conference on Future Networks and Communications (FNC 2017) / Affiliated Workshops, *110*, 386–393. https://doi.org/10.1016/j.procs.2017.06.081

Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence* (Technical Report #2019-1). Future of Humanity Institute, University of Oxford.

Durkheim, É., & Mauss, M. (1963). *Primitive classification* (R. Needham, Trans.). Cohen & West. (Original work published 1903)

Encouraging Earlier Tax Returns in Indonesia. (2018). Behavioural Insights Team. https://www.bi.team/publications/encouraging-earlier-tax-returns-in-indonesia/

Foucault, M. (1977). *Discipline and punish: The birth of the prison* (1st American). Pantheon Books.

Foucault, M. (1982). The Subject and Power. *Critical Inquiry*, *8*(4), 777–795. JSTOR.

Fourcade, M., & Healy, K. (2017). Seeing like a market. *Socio-Economic Review*, *15*(1), 9–29. https://doi.org/10.1093/ser/mww033

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167–194). MIT Press. http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=3339732

Gilliom, J. (2001). *Overseers of the Poor: Surveillance, Resistance, and the Limits of Privacy*. University of Chicago Press.

Hadzic, A., Christie, G., Freeman, J., Dismer, A., Bullard, S., Greiner, A., Jacobs, N., & Mukherjee, R. (2020). Estimating Displaced Populations from Overhead. *ArXiv:2006.14547 [Cs.CV]*.

Hull, M. S. (2012). *Government of Paper* (1st ed.). University of California Press; JSTOR. www.jstor.org/stable/10.1525/j.ctt1pppk9

Introna, L. D. (2015). Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science, Technology, & Human Values*, *41*(1), 17–49. https://doi.org/10.1177/0162243915587360

Jia, C., Yi, W., Wu, Y., Huang, H., Zhang, L., & Wu, L. (2020). Abnormal activity capture from passenger flow of elevator based on unsupervised learning and fine-grained multi-label recognition. *ArXiv:2006.15873 [Cs.CV]*.

Jütte, D. (2014). Entering a city: On a lost early modern practice. *Urban History*, *41*(2), 204–227. Cambridge Core. https://doi.org/10.1017/S096392681300062X

Kerr, I. (2010). Digital Locks and the Automation of Virtue. In M. Geist (Ed.), *From 'Radical Extremism' to 'Balanced Copyright': Canadian Copyright and the Digital Agenda* (p. 247). Irwin Law.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A Conditional Transformer Language Model for Controllable Generation. *ArXiv:1909.05858 [Cs.CL]*.

Latour, B. (1990). Technology is Society Made Durable. *The Sociological Review*, *38*(1_suppl), 103–131. https://doi.org/10.1111/j.1467-954X.1990.tb03350.x

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artefacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 151–180). MIT Press.

Lukes, S. (1974). *Power: A radical view*. Macmillan.

M. Pietrowicz, C. Agurto, J. Casebeer, M. Hasegawa-Johnson, K. Karahalios, & G. Cecchi. (2019). Dimensional Analysis of Laughter in Female Conversational Speech. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6600–6604. https://doi.org/10.1109/ICASSP.2019.8683566

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., & Prabhumoye, S. (2020). Politeness Transfer: A Tag and Generate Approach. *ArXiv:2004.14257 [Cs.CL]*.

Manolev, J., Sullivan, A., & Slee, R. (2019). The datafication of discipline: ClassDojo, surveillance and a performative classroom culture. *Learning, Media and Technology*, *44*(1), 36–51. https://doi.org/10.1080/17439884.2018.1558237

Maronikolakis, A., Villegas, D. S., Preotiuc-Pietro, D., & Aletras, N. (2020). Analyzing Political Parody in Social Media. *ArXiv:2004.13878 [Cs.CL]*.

Milano, S., Taddeo, M., & Floridi, L. (2021). Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, *37*(1), 35–45. https://doi.org/10.1080/01972243.2020.1832636

Nickles, D. P. (2003). *Under the Wire: How the Telegraph Changed Diplomacy*. Harvard University Press. http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=3300605

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Rosenblat, A. (2018). *Uberland* (1st ed.). University of California Press; JSTOR. www.jstor.org/stable/10.1525/j.ctv5cgbm3

Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press; nlebk. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=187883&site=ehost-live&authtype=ip,uid

Shing, H.-C., Resnik, P., & Oard, D. W. (2020). A prioritization model for suicidality risk assessment. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8124–8137.

Smith, I. (2020, March 23). *Creating the COVID-19 text service for vulnerable people*. NHS Digital. https://digital.nhs.uk/blog/transformation-blog/2020/creating-a-covid-19-text-service-for-vulnerable-people

Taylor, A. (n.d.). *An obscure E.U. regulation may have saved lives in the Berlin Christmas market attack*. Washington Post. Retrieved 13 February 2020, from https://www.washingtonpost.com/news/worldviews/wp/2016/12/29/an-obscure-e-u-regulation-may-have-saved-lives-in-the-berlin-christmas-market-attack/

Tilly, C. (1990). *Coercion, capital, and European states, A.D. 990-1990*. Basil Blackwell.

Tilly, C. (2002). *Stories, identities, and political change*. Rowman & Littlefield.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136. JSTOR.

Winner, L. (1986). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press. http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=557593

Woolley, S. C., & Howard, P. N. (2018). Introduction. In S. C. Woolley & P. N. Howard (Eds.), *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press. https://doi.org/10.1093/oso/9780190931407.001.0001

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

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, *12*(4), 505–523. Business Source Complete.