MAX PLANCK INSTITUTE
FOR BEHAVIORAL ECONOMICS

Christoph Engel

Discussion Paper
2026/3

LEGITIMACY IN THE AGE OF ARTIFICIAL INTELLIGENCE

1. Why do we care about legitimacy?

Then he said to them,
"So give back to Caesar what is Caesar's,
and to God what is God's."
Matthew 22:21

Is legitimacy a descriptive or a prescriptive category? The most appropriate response is: it is both. But for conceptual clarity, it is worth distinguishing. The quote from Matthew is prescriptive. The Bible, and the church for that matter, not only acknowledge worldly power. They even stipulate a duty to obey. But it is implicit in the quote that Caesar is not a robber. The duty rests in the implicit assessment that Caesar's power is legitimate.

This is what centuries later Max Weber would characterise as the traditional source of legitimacy. He contrasts it with a rational source (legitimate government solves real social problems) and with charismatic government (legitimate government makes a compelling case) (Weber 1976). Philosophers have aimed at grounding the duty to obey (Dworkin 1986) in actual (Locke 1690) or hypothetical consent (Rawls 2005), in the fact that those in power have acted fairly and have respected reciprocity (Hart 2017) or have provided a true service to the individuals whom they govern (Raz 1986). The duty to obey may also be grounded in deontological principles (Kant 1785) or rule utilitarianism (Brandt 1959), i.e. in the claim that everybody will be better off if everybody obeys the rules that apply to everybody – again premised on these rules being recognised as legitimate.

The most important practical implication of legitimacy as a prescriptive concept is the opposite situation: the duty to obey is contingent on the exercise of sovereign powers being legitimate. Some constitutions, as for instance the German, go as far as stipulating an explicit right to resistance against illegitimate government (Art. 20 IV Basic Law). Even short of this radical legal implication there may be remedies of which citizens can avail themselves if those in power have not acted in a legitimate way. Again German constitutional law is precise about this. Ideally, an act of government (be that an act of legislation, administration or adjudication) should have both: substantive and personal legitimacy. Both should rest in the ultimate source of legitimacy, which is the people. Through elections, the people delegate the exercise of sovereign powers to the legislator. This is why German constitutional law requires that exercises of sovereign powers are grounded in a decision of the legislator. For any interference with freedom or property, German constitutional law requires an explicit delegation in a statute. Moreover, constitutional law requires that sovereign powers be only exercised by public officials who have come into office by a decision that has been taken with the consent of government that has come into office by a decision of Parliament. In constitutional practice, to a degree these strict requirements may be qualified. It may, for instance, be acceptable that private entities exercise delegated powers. But such exceptions require explicit justification, for instance by superior expertise of the private entity (for background and references see Seiler 2025).

One elaborate version of legitimacy as a descriptive concept has been developed in political science. It is the distinction between input and output legitimacy (Scharpf 2009). Input legitimacy means that sovereign powers are exercised by agents who have been endowed with these powers by the supreme authority; in a democracy, this authority is the electorate. Hence in a democracy, input legitimacy is tied to the concept of participation (Verba 1967). Input legitimacy becomes questionable if those originally elected try to sidestep democratic control, and if the regime is on the slippery slope towards autocracy (Gerschewski 2018). To remain legitimate, a democratic government must be open to change, including the possibility that the ruling party loses office on the next election. Another obvious reason to doubt input legitimacy is the suspicion of corruption (Seligson 2002).

The companion concept is output legitimacy. The concept has specifically been developed for the (original setup of) the European Union. As the European Union is not a full-blown state, its powers have always been a hybrid between direct responsibility towards the people (through the European Parliament) and indirect responsibility through the governments of the member states. This hybrid make-up has, in political debates, being used to delegitimise the European Commission and the European Council. The concept of output legitimacy is meant as a counter-argument. It argues that traditional input legitimacy is not the only possible source. Legitimacy may also result from successful problem-solving, like the establishment of a common market, or the fulfilment of widely acknowledged social or environmental goals (Scharpf 2009). In the later debate, throughput legitimacy has been added as a third category (Schmidt 2013). This concept stresses the procedure of generating and applying rules. It can be related to the claim from system theory that, in complex modern societies, legitimacy can only result from the way how social issues are dealt with (Luhmann 1969).

As the concept is descriptive, it may be put to the test. In a lab experiment, we have compared the force of input and output legitimacy. Participants had to earn a fixed amount of income with a tedious real effort task. They were supposed to pay 30% tax, to be spent on a donation to one of 10 organisations on a list. The list was deliberately constructed such that most student participants would likely support some, but not all recipients. The treatment manipulation was the option to pay a small amount of the earned income for switching from a regime in which the recipient was randomly selected from the list to a varied degree of influence on this selection. There were three regimes. In the first regime, the majority recipient would get all the taxes. In the second regime, the three most popular regimes would get some of the amount. In the third regime, the more votes one recipient had collected the greater the share from the entire tax income. Participants were sensitive to the degree of input legitimacy. The more personal influence they had, the more of their income they were willing to spend on being in that regime. Moreover the bigger their individual influence, the more they were tax abiding. But there was a catch. We had elicited beliefs about the most popular recipients. It turned out that most participants were (correctly) believing that recipients they deemed acceptable themselves would also be selected by others. We concluded from this result that, actually, what looked like strong support for input legitimacy effectively was an attempt at channelling the money where they wanted it to be spent. Or using the terminology from political science: participants sought out input legitimacy as a technology to achieve output legitimacy (Engel, Mittone et al. 2024).

In an experiment, it is possible to isolate true input legitimacy and true output legitimacy. In political reality, this is difficult. Whether a person has actual influence on the selection of members of Parliament and, through them, on choices made by elected politicians, is something the individual voter cannot truly verify. Even if the election does not appear rigged, the channel of influence is highly indirect. Even more indirect is the influence of voters on individual substantive political decisions. In countries like Switzerland, there is more scope for direct democracy through referendums. But even in such countries, direct democratic decision-making is the exception that proves the rule of the representation principle. For these reasons, in political practice, descriptive legitimacy is essentially a matter of beliefs. Whether citizens regard an exercise of sovereign powers as legitimate is filtered through perception (Van der Toorn, Tyler et al. 2011, Varet, Granié et al. 2021). Essentially it is a matter of trust (Hawdon 2008, Hough, Jackson et al. 2013, Batrancea, Nichita et al. 2019). In general, trust is easy to destroy, but hard to build (Chaudhuri and Gangadharan 2007, Charness, Du et al. 2011). Trust in the legitimate exercise of sovereign powers is no exception to this rule. Trust relies on signals (Spence 1973). This question features centrally in research that tries to isolate the reasons why most (legal) rules are applied even if the addressee has no reason to suspect that rule violation would lead to audit or enforcement (Desmet and Engel 2021). This research demonstrates that the most important source of compliance, with almost any legal rule, is the way how this individual feels treated when she, or someone from her close personal environment, has been treated by the administration or the courts (Tyler 2006b, Tyler 2006a). Transparency helps, in that it increases the information base. But it may also hurt, in that occasional mistakes become widely known (Engel 2019).

2. How could artificial intelligence affect legitimacy?

In a nutshell, this paper asks: is legitimacy technology neutral? Is everything that has changed the mechanical way in which sovereign powers are gained or exercised? Or is sovereignty qualitatively different if it is grounded in AI? Sovereignty has never been technology free. Once the printing press had been invented, it has been used for mustering resistance to those in power. The radio has helped mobilize the electorate. E-mail campaigns have deprived traditional media of their gate keeping power. Those designing new rules have consulted libraries. They have collected data, for instance about the number of taxpayers, or about the quality of the infrastructure. They have run surveys to learn more about citizens' preferences. Those implementing governmental rules have kept records about past governmental acts, and have used them to predict the success of current interventions. For regulating street traffic, machines (traffic lights) have long replaced policemen. The tax administration uses software to target audits to the most suspicious cases. A further reminder is in place: new technologies that have been used to gain or exercise sovereign powers have rarely been invented or produced by government, for the purpose of governing society. Gutenberg has invented the printing press to make some money. The radio has originally been built for mass entertainment. E-mail was designed as a technology to survive after a nuclear strike. Those combatting for sovereign power, and governments exercising it, have availed themselves of the new opportunities that had originally been developed outside government, and for non-governmental purposes.

This invites a more precise question: in which ways does artificial intelligence provide new opportunities for legitimate government, and new challenges? This question may be asked at four levels: access to political power, rule generation, rule application, and shaping up support for government.

a). Access to political power

In a democracy, political power is precarious by design. The very fact that those who intend to wield political power have to fight for it, and that they know they have to survive the next election, is itself a source of legitimacy. Now elections have never been mere exercises in preference aggregation. In a representative democracy, voters make a highly aggregated decision. They have to select an individual whom they, everything considered, deem acceptable. In modern democracies, the affiliation of a candidate with a political party is the most important piece of information. This mediatisation of candidates by large organisations further coarsens the choice of a voter. Moreover in a typical election, the influence of the individual voter is minimal, which is why political scientists have coined the term “voting paradox”. It stresses that an individual voter is next to never pivotal (Ferejohn and Fiorina 1974).

The big change brought about by artificial intelligence is a drastic reduction in transaction cost (on transaction cost economics, see Williamson 1975). In the past, the party systems of most democracies have been very stable over time. New parties, like the Greens, have made it into the system. But such changes have taken decades. With the help of artificial intelligence, mobilising individuals and organising them, has become dramatically cheaper. In the language of competition theory, barriers to entry into the political market are substantially lower (von Weizsäcker 1980, Baumol and Willig 1981). This creates a new opportunity for groups of citizens who did not feel adequately represented by the existing party system. Access to political power is democratized. But at the same time, the party system becomes more fluid. Political stability is harder to establish.

In the political theory textbook, the individual voter compares the political programs of the competing parties and selects the one that comes closest to her preferences. Were she to do that by the rulebook, she would have to assign a normative weight to each item that she expects to be on the agenda for the upcoming legislative period. She would gauge the gap between the political decision she desires herself and the position of the party. She would also have to factor in the probability that the decision will be made at all, and the need for compromises between political parties if the country will likely be ruled by a coalition. Empirically, this is of course not what the typical voter does. Rather she goes by a few salient issues about which she cares most, she has been convinced by an individual candidate, or she votes again for the same party for which she has voted earlier. She may even refrain from exercising her voting right altogether.

Again artificial intelligence creates new opportunities. If a voter dearly cares about an issue, she may ask a language model about the expressed or inferred preferences of the competing parties. She may also ask the model to develop scenarios that would be required for her desired outcome to obtain. With the help of AI, a voter who truly cares may not only be better informed. She also stands a more realistic chance to engage other voters about what she considers to be a critical issue. The voting paradox only holds for the isolated decision in the voting booth, not for groups of voters who generate a movement prior to the election day.

Those competing for a seat in parliament may avail themselves of AI technology in the interest of attracting votes. Again, the normative assessment is ambivalent. Political parties, and their candidates, must address a very heterogeneous electorate. Not every voter cares about the same issue. Artificial intelligence makes it possible to address different voters differently. Not only has such a more personalised campaign become technically and economically feasible (Enli and Skogerbø 2013). Candidates may also exploit demographic markers to send the right message to the right person. Per se, this is not reproachable. But there is of course the possibility to hide part of the intended program to voters the candidate expects to be less inclined towards such issues. The dividing line between personalisation and manipulation may be narrow.

Candidates, and entire political parties for that matter, must excel in the art of meeting competing, if not conflicting wishes about the evolution of the polity (Brunsson 1989). But even taking this into account, they must sometimes make choices. Either the campaign stresses an issue, or it plays it down. The better the information about the distribution of wishes in the population, the more it is promising for the campaign to exploit the power of large language models to handle rich data. The campaign managers may exploit this ability to simulate voter turn-out and votes conditional on alternative versions of the campaign (Aher, Arriaga et al. 2023, Horton 2023). To make these predictions as informative as possible, the campaign may build a set of digital twins of the population (Bagabaldo and Hackl 2025, Helm, Chen et al. 2025).

In democracy, an election may be challenged if there has been a glaring manipulation. But most constitutions make this difficult. The legal and practical impediments result from a balance between deterring violations of electoral law, and respecting the sense of justice of those who feel treated unfairly on the one hand, and the dangers of political instability on the other hand. Using economic language, one may also say that the official outcome of the election has changed the political equilibrium. A deontological definition of legitimacy, or one derived from rule utilitarianism, could support this line of reasoning. As long as the underlying rules of electoral law are in line with the constitution, there would also not be a problem with the legal definition of legitimacy. But the serious suspicion that, with the help of artificial intelligence, the election has been rigged would seriously taint legitimacy as a political category. Input legitimacy would be qualified. Same for philosophical concepts that ground the duty to comply in imputed consent, fairness or reciprocity.

b) Rule generation

Max Weber's "rational" source of legitimacy results from the ability of government to solve social problems. This is also the core of the concept of input legitimacy in political science (Simon 1986). Some problems of a country require an ad hoc intervention. If there is a crisis, it must be addressed, even if the crisis has not been anticipated, so that there is no legislative rulebook. But many social problems are of a different kind. The risk of crime is foreseeable, as is climate change, or the problems with a pay as you go pension system if the age distribution in the population is unbalanced. With foreseeable social problems, the distinction between rule generation and rule application is meaningful. Artificial intelligence changes both.

Some social problems are simple. But even seemingly simple problems, like the prohibition of stealing, may become more complicated when looking more closely. To stick with the example: the typical thief does not steal for a living (cf. Levitt and Venkatesh 2000). Even those who do did not necessarily chose this profession voluntarily over alternative careers. And even if they did, a former thief may not have easy access to the official labour market (Agan and Starr 2018). Consequently, the political goal of fixing the stealing rate cannot necessarily be achieved with the recipe from rational choice theory, i.e. deterrence (Nagin 1998). Different countries have tried out different solutions, like a more encompassing welfare state. But alas, even in the Nordic countries with their strong welfare state some stealing happens (Apel 2025).

In similar ways as with attracting voters, artificial intelligence can make political interventions more effective that aim at solving a defined social problem. Government may access richer information. It can simulate the effects of alternative interventions. It may address heterogeneous populations by conditioning interventions on discernible demographic markers. If the social problem is likely to persist for an extended period of time, government may evaluate the effect of a first intervention, and adjust it in the light of fine-grained information about its implementation. This way, rule generation can be turned into a dynamic process which builds on social learning (Bennear and Wiener 2019).

There are a number of downsides though. A widely discussed risk are hallucinations (Dahl, Magesh et al. 2024, Magesh, Surani et al. 2025). The risk results from the architecture of large language models. Under the hood, they are prediction engines. This also holds for their generative ability. Technically the next token (typically the next word) is generated as a prediction based on the generated text up till the last word. The problem is exacerbated by a design feature of most language models. They are programmed to give a response, even if they are only mildly confident that the response is correct. While the problem is serious, for rule generation it is probably not too concerning. Hallucinations are often obvious to a human observer. Rule generation is unlikely to be fully automated. When checking back, human supervisors are likely to spot the mistake.

This is a very different with another feature of large language models. Recent frontier models are trained on huge amounts of data, essentially the part of the Internet that is not behind paywalls. For this reason, large language models have typically not only read more human utterances than any human individual will read in their lifetime. Different large language models are also likely to have been trained on similar data (for an overview see Bhati, Neha et al. 2026). But the design of large language models is not confined to exploiting the information inherent in the training data. Different LLMs use very different approaches for making the most out of this information. For that purpose they use reinforcement with human feedback (Ouyang, Wu et al. 2022), and this feedback tends to be proprietary. To the degree possible, they also use reinforcement with verifiable rewards, provided the clear and automatic evaluation of the output is feasible (as epitomised by mathematical or coding tasks) (Wen, Liu et al. 2025). It has become increasingly popular to supplement scraped real data with synthetic data (Adler, Agarwal et al. 2024). Finally, the providers implement different "guardrails", for instance to prevent the large language model to be used for crime, or for other purposes that the provider deems inappropriate (Rebedea, Dinu et al. 2023).

Importantly, these interventions endow different large language models with pronouncedly different normative convictions. In an experiment, we have tested four state of the art large language models on three different versions of the trolley dilemma (Foot 1967). In the bystander problem, an observer notices that a trolley would run into five people working on a track. By pulling a lever, the observer could redirect the trolley to another track, but the track would then kill another person. Should she do that? In the footbridge variant, instead of pulling a lever, the observer would have to push a heavy person over the bridge such that this person would stop the trolley, and die (Thomson 1984). In the triage variant, the choice is one between saving five patients at the expense of another patient who would sure survive otherwise (Christian 2019). We have tested the four LLMs of Figure 1 on all three problems. We also happened to have data from human participants (Engel, Hermstrüwer et al. 2025).

The blue bars stand for the percentage of intervention decisions (pulling the lever). In the bystander problem, all four LLMs agree, but they are all more interventionist than human controls. In the remaining two problems, there is a stark difference between Claude and the remaining LLMs. While Claude is very hesitant to intervene, the other three LLMs are quite happy to sacrifice one life for many.

The idiosyncratic and undisclosed normative stance of an LLM would be less concerning if the human legislator could instruct the LLM to reason along the lines it considers normatively adequate. This we have tested with three interventions. In the first intervention, we tell the LLM to follow the deontological logic. In the third intervention, we adopt the opposite perspective and instruct it to decide in line with utilitarian principles. The second, in between intervention instructs the model to strike a balance between both normative perspectives. There is some good news. Gemini is quite responsive to either instruction. Mistral also exhibits some reaction. Llama is, however, quite unswerving. And Claude is not easy to convince that a decision could be taken on utilitarian grounds. Our results suggest that "human realignment" is not an impossible endeavour. The experiment also shows that the hidden normative position of an LLM can be revealed with the help of experiments. But legislators should not expect an LLM to be normatively neutral if they use it for the preparation of new rules.

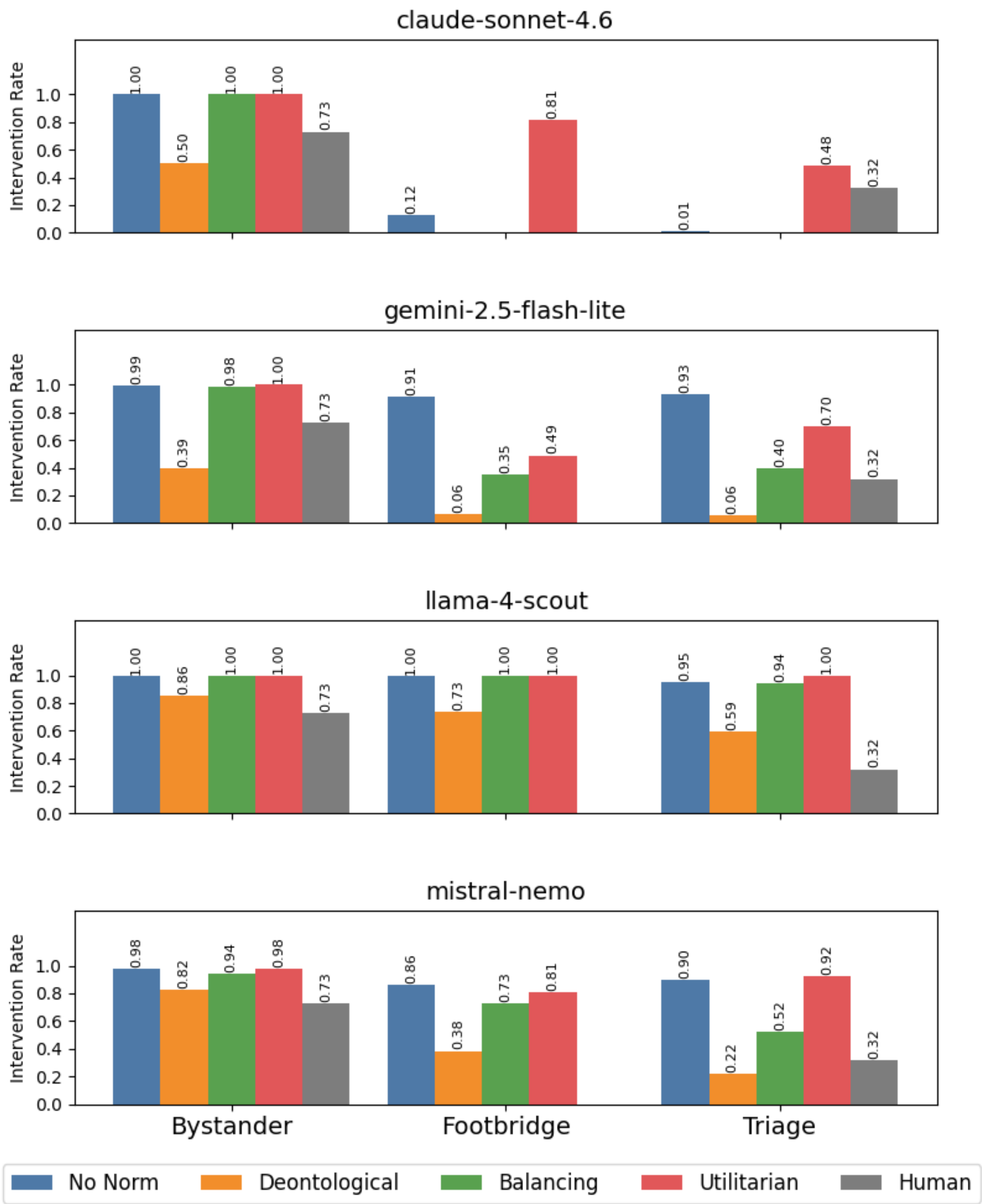

Figure 1
Frequency of sacrificing one person for the sake of five others

A related concern is the proprietary nature of all frontier models. All of them have been developed by business. The greater their impact on rule generation, the more legislative powers would shift to private entities. Political decisions might be transformed into technocratic choices. Moreover Mistral is the exception that proves the rule: all other serious competitors either come from the US or from China. It is not known to which degree their governments have influenced the design, and control their deployment. At any rate, this creates a degree of vulnerability of unknown severity.

The price for frontier models to be highly effective is opacity (Burrell 2016). Some first generation machine learning algorithms (like decision trees) were transparent and easily explainable. This benefit already disappeared with more powerful machine learning tools, like support vector machines. With the next generation, the architecture shifted towards neural networks with hidden layers. Even if a supervisor checked those hidden layers, she would know very little about their impact on the ultimate prediction. Opacity further increased with the translation of verbal input into high-dimensional vectors of probabilities, so-called embeddings. The generative component of large language models made the internal workings of large language models completely obscure (an excellent introduction into the technical development is James, Witten et al. 2022). Effectively, the only way to learn something about the way how a frontier model works is running experiments. From appropriately designed interventions one may infer some information about the relationship between input and output. Of course, this is no different with human decision makers. The inner workings of the human brain are at least as intransparent as are language models. Yet human decisionmakers have thousands of years of experience to interact with other human decision makers, while language models are so recent that no such social practice has emerged.

Even assuming a large language model that is fully under national democratic control, the mere ability of designing much more specific rules creates an issue for legitimacy. The technology has the potential of shifting decision-making power away from administrators who apply more generic rules to rule makers who design more fine-grained interventions (Langenbach 2026).

The true power of large language models as decision aids for legislators rests in their ability to aptly handle huge amounts of information. This is why large language models may help the legislator design more effective interventions. But this advantage requires that the legislator has access to such fine-grained information, and makes it available to the large language model. Whenever the addressee of the planned intervention is an individual, knowing more about this specific addressee makes intervention smarter. But inevitably government either needs to collect this information in the first place, or it must take this information from a data collection enterprise that has been run for different purposes. The most practical source is the data that private entities have collected anyways for their commercial purposes. Government might either buy that data, or might even conscript it. Yet the more private data government plans to use in preparation of a new rule, the more deeply it interferes with privacy. Moreover, citizens who are wary of the use of personal information for regulatory purposes may try to conceal this information, in the interest of thwarting the planned intervention.

c) Rule application

Similar considerations hold if the administration (or the courts) want to apply the rules that the legislator has designed to individual cases. There is considerable potential in mustering the power of large language models for the purpose. But there is also a list of concerns. Consequently, when relying on LLMs, legitimacy can increase because government exhibits greater problem-solving capacity. But legitimacy can also decrease because the way how government becomes more effective is considered inappropriate.

The implementation deficit is a pervasive problem of the modern state (Mayntz 1983, Knill, Steinebach et al. 2024). Rules that serve a fully legitimate, uncontested purpose partly or even fully fail to achieve their intended purpose because government cannot implement them. Often the implementation deficit does not hold across the board, but affects some parts of the population more severely than others. If that is the case, the normative concern is exacerbated by discrimination between discernible groups of society. Because many are getting away with rule violations, the willingness of good natured citizens to abide by the rule erodes (for an experimental test of the effect see Desmet and Engel 2021). If this is not an exceptional event, general support for government may even be at risk.

Large language models have the potential to make rule application much more effective. The main reason is again transaction cost. Once the relevant part of social life leaves digital traces, the cost of audit is drastically reduced. Implementation may be near perfect, for instance by only giving access to the desired activity after the LLM has received proof of compliance with the applicable rules. If government does not want to go that far, it may keep track of prior interactions with a citizen, and target audit towards individuals that the LLM predicts to be more likely to violate the rule. If citizens try to circumvent a rule, the LLM may flag attempted circumvention for audit.

LLMs can exploit more information about the concrete situation and the individual citizen whose behaviour the administration wants to affect. The resulting greater precision of the intervention does not only make the individual intervention more effective. The administration may also better target the situation and the addressee which promise the greatest effect. Through customisation, individual interventions can be less brushing. For most addresses this may mean a less intrusive approach, while implementation efforts may zero in on the situations and individuals who are most relevant for achieving the regulatory goal (Ben-Shahar and Porat 2021).

While these are serious opportunities for increasing legitimacy, essentially all the potential concerns that have been spelt out with rule generation do also affect rule application. There is an additional challenge which is much more important for rule application than for rule generation. If the administration or the courts want to fully exploit the power of large language models, they must shift to a fully digital workflow. The available information about the case and the addressee are not only processed with the help of a large language model. These informations are themselves digitized (and made available in a standardized, easily machine-readable format, like .json). As the addressees know that their personal interests are at stake, they have a strong incentive to manipulate the advice or decision given by the LLM. This is concerning because LLMs can be manipulated in ways that are not immediately transparent to human users, and to a supervising administrator or judge in particular. One primitive, but still not yet fully fixed risk is prompt injection (Gulyamov, Gulyamov et al. 2026). The addressee for instance writes an instruction in white font on white background that the LLM easily reads, while the human supervisor will not notice it. In this instruction, the addressee request from the LLM to decide in her favour. Hence from the perspective of the legitimacy balance, the involvement of artificial intelligence is a double-edged sword.

d) Perception

If the substantive or personal legitimacy of a statute, administrative act or court ruling is challenged in the constitutional court, the court may investigate the matter thoroughly. But the main reason for discussing legitimacy is not the legality of the exercise of sovereign powers. The critical question is whether government is considered legitimate by the population. This is why perception is important.

If the country, or the way how it is governed, are perceived to be in good shape, this may convey government legitimacy. Problem-solving capacity matters for perceived legitimacy. But there is no one-to-one mapping between the problem-solving performance of government and its perceived legitimacy. One straightforward reason is an information gap. The wider public may simply not know about actual good governance, or it may not have noticed bad governance (Achen and Bartels 2017). The characteristic short-term orientation of democracy to the next election day also matters in this respect (Ogami 2024). Moreover democratic government is not exclusively about solving serious social problems. Typically there is more than one solution, and these solutions differ by their distributional effects. One solution favours one group, while another solution favours another group (Knight 1992). Such constellations are particularly dangerous if the group that could easiest solve the problem is also best organised (Olson 1965). Then the better organised group is likely to exert pressure on government, and government might try to conceal from the wider public that it has given in to the pressure.

The involvement of artificial intelligence in the generation and the application of sovereign rules may affect perceived legitimacy of government in a generic and in a specific way. The generic channel is the attitude of the public towards artificial intelligence. First generation research has documented pronounced "algorithm aversion" (Dietvorst, Simmons et al. 2015). Putting a "human in the loop" is the classic response (Odekerken, Bex et al. 2024). But more recent evidence is more nuanced. It shows that attitudes towards AI tend to be context specific (Starke, Baleis et al. 2022, Hermstrüwer and Langenbach 2023), and can be changed by targeted design (Henning and Langenbach 2026).

Since for the most part legitimacy is constructed through the lens of perception, perception management is a classic governmental task (Carpenter and Krause 2012, Guenther, Jörges et al. 2024, Sunstein and Gaffe 2024). Government may muster the power of AI for this persuasion task (Cheng and You 2025), in particular by personalising information (Xi, Zeng et al. 2026).

3. Legitimacy in the age of artificial intelligence

The key argument of this paper is: technology shapes how sovereign powers may be exercised, and how those respond who anticipate the exercise of sovereign powers. Technology creates opportunities for governing society, and it creates challenges for effective governance. Technology thus affects the problem-solving capacity of those in power, and it affects how sovereign power is perceived. This is why legitimacy is not technology neutral. Artificial intelligence fundamentally facilitates the ability to process information.

Whether this change in abilities is good or bad for the legitimacy of sovereign powers is ambivalent. Neither does the use of artificial intelligence make the acquisition and the exercise of sovereign powers illegitimate. Nor does the involvement of the new opportunities created by big data and artificial intelligence necessarily make the exercise of sovereignty, or the way how it is perceived by the public, more legitimate. The bottom line is the classic response of jurists: it depends. This paper has sketched some of the channels through which the deployment of artificial intelligence is likely to influence actual and perceived legitimacy of government. The balance between new opportunities and new risks will depend on how governments use the new opportunities, how they contain the new risks, and how this is communicated to the public.

References

Achen, Christopher H and Larry M Bartels (2017). Democracy for Realists. Why Elections Do Not Produce Responsive Government, Princeton University Press.

Adler, Bo, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay and Jonathan Cohen (2024). “Nemotron-4 340b Technical Report.” https://arxiv.org/pdf/2406.11704.

Agan, Amanda and Sonja Starr (2018). "Ban the Box, Criminal Records, and Racial Discrimination. A Field Experiment." Quarterly Journal of Economics **133**(1): 191-235.

Aher, Gati V, Rosa I Arriaga and Adam Tauman Kalai (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. International Conference on Machine Learning, PMLR.

Apel, Robert (2025). "Can Social Safety Net Spending Prevent Crime?" Annual Review of Criminology **9**: 101-125.

Bagabaldo, Alben Rome and Jürgen Hackl (2025). “Digital Twins for Intelligent Intersections: A Literature Review.” https://arxiv.org/pdf/2510.05374.

Batrancea, Larissa, Anca Nichita, Jerome Olsen, Christoph Kogler, Erich Kirchler, Erik Hoelzl, Avi Weiss, Benno Torgler, Jonas Fooken and Joanne Fuller (2019). "Trust and Power as Determinants of Tax Compliance across 44 Nations." Journal of Economic Psychology **74**: 102191.

Baumol, William J. and Robert D. Willig (1981). "Fixed Cost, Sunk Cost, Entry Barriers and Sustainability of Monopoly." Quarterly Journal of Economics **95**: 405-431.

Ben-Shahar, Omri and Ariel Porat (2021). Personalized Law. Different Rules for Different People, Oxford University Press.

Bennear, Lori S and Jonathan B Wiener (2019). “Adaptive Regulation: Instrument Choice for Policy Learning over Time.” https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/Regulation%20-%20adaptive%20reg%20-%20Bennear%20Wiener%20on%20Adaptive%20Reg%20Instrum%20Choice%202019%2002%2012%20clean.pdf.

Bhati, Deepshikha, Fnu Neha, Devi Sri Bandaru, Matthew Weber and Ishan Dilipbhai Gajera (2026). “Large Language Models. A Survey of Architectures, Training Paradigms, and Alignment Methods.” https://www.preprints.org/frontend/manuscript/abaf3e6f650e2aeab454da56761a79aa/download_pub.

Brandt, Richard B (1959). Ethical Theory.

Brunsson, Nils (1989). The Organization of Hypocrisy. Talk, Decisions, and Actions in Organizations. Chichester ; New York, Wiley.

Burrell, Jenna (2016). "How the Machine ‘Thinks’. Understanding Opacity in Machine Learning Algorithms." Big Data & Society **3**(1): 2053951715622512.

Carpenter, Daniel P and George A Krause (2012). "Reputation and Public Administration." Public Administration Review **72**(1): 26-32.

Charness, Gary, Ninghua Du and Chun-Lei Yang (2011). "Trust and Trustworthiness. Reputations in an Investment Game." Games and Economic Behavior **72**(2): 361-375.

Chaudhuri, Ananish and Lata Gangadharan (2007). "An Experimental Analysis of Trust and Trustworthiness." Southern Economic Journal **73**: 959-985.

Cheng, Zirui and Jiaxuan You (2025). “Towards Strategic Persuasion with Language Models.” https://arxiv.org/pdf/2509.22989.

Christian, Michael D (2019). "Triage." Critical Care Clinics **35**(4): 575-589.

Dahl, Matthew, Varun Magesh, Mirac Suzgun and Daniel E. Ho (2024). "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." arXiv preprint arXiv:2401.01301.

Desmet, Pieter and Christoph Engel (2021). "People Are Conditional Rule Followers." Journal of Economic Psychology **85**: 102384.

Dietvorst, Berkeley J, Joseph P Simmons and Cade Massey (2015). "Algorithm Aversion. People Erroneously Avoid Algorithms after Seeing Them Err." Journal of Experimental Psychology: General **144**(1): 114-126.

Dworkin, R. M. (1986). Law's Empire. Cambridge, Mass., Belknap Press.

Engel, Christoph (2019). "When Does Transparency Backfire? Putting Jeremy Bentham's Theory of General Prevention to the Experimental Test." Journal of Empirical Legal Studies **16**(4): 881-908.

Engel, Christoph, Yoan Hermstrüwer and Alison Kim (2025). “Human Realignment. An Empirical Study of Llms as Legal Decision-Aids in Moral Dilemmas.” https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5216030.

Engel, Christoph, Luigi Mittone and Azzurra Morreale (2024). "Outcomes or Participation? Experimentally Testing Competing Sources of Legitimacy for Taxation." Economic Inquiry **62**: 563-583.

Enli, Gunn Sara and Eli Skogerbø (2013). "Personalized Campaigns in Party-Centred Politics. Twitter and Facebook as Arenas for Political Communication." Information, Communication & Society **16**(5): 757-774.

Ferejohn, John A. and Morris P. Fiorina (1974). "The Paradox of Not Voting. A Decision Theoretic Analysis." American Political Science Review **68**: 525-536.

Foot, Philippa (1967). "The Problem of Abortion and the Doctrine of Double Effect." Oxford Review **5**: 5-15.

Gerschewski, Johannes (2018). "Legitimacy in Autocracies. Oxymoron or Essential Feature?" Perspectives on Politics **16**(3): 652-665.

Guenther, Lars, Susan Jörges, Daniela Mahl and Michael Brüggemann (2024). "Framing as a Bridging Concept for Climate Change Communication. A Systematic Review Based on 25 Years of Literature." Communication Research **51**(4): 367-391.

Gulyamov, Saidakhror, Said Gulyamov, Andrey Rodionov, Rustam Khursanov, Kambariddin Mekhmonov, Djakhongir Babaev and Akmaljon Rakhimjonov (2026). "Prompt Injection Attacks in Large Language Models and Ai Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms." Information **17**(1): 54.

Hart, Herbert LA (2017). Are There Any Natural Rights? Theories of Rights, Routledge**:** 61-77.

Hawdon, James (2008). "Legitimacy, Trust, Social Capital, and Policing Styles. A Theoretical Statement." Police Quarterly **11**(2): 182-201.

Helm, Hayden, Tianyi Chen, Harvey McGuinness, Paige Lee, Brandon Duderstadt and Carey E Priebe (2025). “Toward a Digital Twin of Us Congress.” https://arxiv.org/pdf/2505.00006.

Henning, Arian and Pascal Langenbach (2026). "Bridging the Human–Ai Fairness Gap. How Providing Reasons Enhances the Perceived Fairness of Public Decision-Making." Journal of Empirical Legal Studies **23**(1): 39-59.

Hermstrüwer, Yoan and Pascal Langenbach (2023). "Fair Governance with Humans and Machines." Psychology, Public Policy, and Law **29**: 525-548.

Horton, John J (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, National Bureau of Economic Research.

Hough, Mike, Jonathan Jackson and Ben Bradford (2013). “Legitimacy, Trust and Compliance: An Empirical Test of Procedural Justice Theory Using the European Social Survey.” https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=2234339.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2022). An Introduction to Statistical Learning, Springer.

Kant, Immanuel (1785). Grundlegung Zur Metaphysik Der Sitten. Riga.

Knight, Jack (1992). Institutions and Social Conflict. Cambridge England, Cambridge University Press.

Knill, Christoph, Yves Steinebach and Dionys Zink (2024). "How Policy Growth Affects Policy Implementation. Bureaucratic Overload and Policy Triage." Journal of European Public Policy **31**(2): 324-351.

Langenbach, Pascal (2026). Die Verwaltung Von Heterogenität. Personalisierung Und Algorithmisierung Verwaltungsrechtlicher Verhaltenssteuerung.

Levitt, Steven D. and Sudhir Alladi Venkatesh (2000). "An Economic Analysis of a Drug-Selling Gang's Finances." Quarterly Journal of Economics **115**(3): 755-789.

Locke, John (1690). Two Treatises of Government. London,, A. Churchill.

Luhmann, Niklas (1969). Legitimation Durch Verfahren. Neuwied am Rhein, Luchterhand.

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning and Daniel E Ho (2025). "Hallucination-Free? Assessing the Reliability of Leading Ai Legal Research Tools." Journal of Empirical Legal Studies **22**(2): 216-242.

Mayntz, Renate (1983). Implementation Von Regulativer Politik. Implementation Politischer Programme Ii. R. Mayntz. Opladen, Westdeutscher Verlag**:** 50-74.

Nagin, Daniel (1998). Deterrence and Incapacitation. The Handbook of Crime and Punishment. M. Tonry. New York, Oxford University Press**:** 345-368.

Odekerken, Daphne, Floris Bex and Henry Prakken (2024). "Precedent-Based Reasoning with Incomplete Information for Human-in-the-Loop Decision Support." Artificial Intelligence and Law **32**: 1-46.

Ogami, Masakazu (2024). "The Conditionality of Political Short-Termism. A Review of Empirical and Experimental Studies." Politics and Governance **12**.

Olson, Mancur (1965). The Logic of Collective Action. Public Goods and the Theory of Groups. Cambridge, Mass., Harvard University Press.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama and Alex Ray (2022). "Training Language Models to Follow Instructions with Human Feedback." Advances in neural information processing systems **35**: 27730-27744.

Rawls, John (2005). Political Liberalism, Columbia University Press.

Raz, Joseph (1986). The Morality of Freedom, Clarendon Press.

Rebedea, Traian, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien and Jonathan Cohen (2023). Nemo Guardrails: A Toolkit for Controllable and Safe Llm Applications with Programmable Rails. Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations.

Scharpf, Fritz Wilhelm (2009). "Legitimacy in the Multilevel European Polity." European Political Science Review **1**: 173-204.

Schmidt, Vivien A (2013). "Democracy and Legitimacy in the European Union Revisited. Input, Output and 'Throughput'." Political Studies **61**(1): 2-22.

Seiler, Christian (2025). Demokratische Legitimation. Handbuch Des Staatsrechts Iii Demokratie. U. Kischel and H. Kube. Heidelberg, C.F.Müller**:** 59-102.

Seligson, Mitchell A (2002). "The Impact of Corruption on Regime Legitimacy. A Comparative Study of Four Latin American Countries." Journal of Politics **64**(2): 408-433.

Simon, Herbert A. (1986). Decision Making and Problem Solving. Report of the Research Briefing Panel on Decision Making and Problem Solving. N. A. o. Sciences. Washington, National Academy Press.

Spence, Michael (1973). "Job Market Signalling." Quarterly Journal of Economics **87**: 355-374.

Starke, Christopher, Janine Baleis, Birte Keller and Frank Marcinkowski (2022). "Fairness Perceptions of Algorithmic Decision-Making. A Systematic Review of the Empirical Literature." Big Data & Society **9**(2): 20539517221115189.

Sunstein, Cass R and Jared H Gaffe (2024). "An Anatomy of Algorithm Aversion." Columbia Science and Technology Law Review **26**: 290-316.

Thomson, Judith Jarvis (1984). "The Trolley Problem." Yale Law Journal **94**: 1395-1415.

Tyler, Tom R (2006a). "Psychological Perspectives on Legitimacy and Legitimation." Annual Review of Psychology **57**: 375-400.

Tyler, Tom R. (2006b). Why People Obey the Law. New Haven, Yale University Press.

Van der Toorn, Jojanneke, Tom R Tyler and John T Jost (2011). "More Than Fair. Outcome Dependence, System Justification, and the Perceived Legitimacy of Authority Figures." Journal of Experimental Social Psychology **47**(1): 127-138.

Varet, Florent, Marie-Axelle Granié, Laurent Carnis, Frédéric Martinez, Marie Pelé and Anthony Piermattéo (2021). "The Role of Perceived Legitimacy in Understanding Traffic Rule Compliance. A Scoping Review." Accident Analysis & Prevention **159**: 106299.

Verba, Sidney (1967). "Democratic Participation." Annals of the American Academy of Political and Social Science **373**(1): 53-78.

von Weizsäcker, Carl Christian (1980). Barriers to Entry. A Theoretical Treatment. Berlin ; New York, Springer-Verlag.

Weber, Max (1976). Wirtschaft Und Gesellschaft. Grundriss Der Verstehenden Soziologie. Tübingen, Mohr.

Wen, Xumeng, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li and Ziming Miao (2025). “Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base Llms.” https://arxiv.org/pdf/2506.14245.

Williamson, Oliver E. (1975). Markets and Hierarchies. Analysis and Antitrust Implications. New York, Free Press.

Xi, Qi, Jing Zeng, Zhanghao Li and Mike S Schäfer (2026). "Personalized Persuasion through Conversational Ai. Can Deepseek Change Perceptions of Genetically Modified Foods in China?" Media and Communication **14**: 11451.