Societal Backfire Risks From Advances in Commitment Technology

Duncan McClements, Urvi Gaur & Martín Soto*

September 2024

Abstract

We present the first comprehensive analysis of potential negative externalities, or "backfire risks", arising from hypothetical future advancements in commitment technology. Through an interdisciplinary review and conceptual analysis, we offer a novel classification of disjunctive risk scenarios, emphasizing the catalytic role that AI development could play in accelerating these risks. Our analysis encompasses a range of potential hazards, from the exacerbation of problems with historical precedent to the unlocking of new possibilities for catastrophic failure. By cataloging historical precedents and analyzing potential future implications, we provide speculative insights and tentative recommendations aimed at mitigating these risks, contributing to the growing discourse on technological governance. But our main recommendation remains continuing the study of these uncertain risks.

1 Introduction

Commitment technologies have played a crucial role in shaping human cooperation and conflict throughout history. From verbal promises to written contracts, and from legal systems to international treaties, these mechanisms have enabled individuals, organizations, and nations to make credible pledges about their future actions. As technology continues to advance in cryptography, artificial intelligence, and distributed systems, we might stand on the brink of a new era in commitment capabilities. This could be precipitated by enhanced transparency between institutions and processes, increasingly pervasive arbitration of social or private spheres, cheaply automated enforcement of contracts, or the development of more sophisticated reputational systems, among others.

While the past and potential benefits of stronger commitment technologies are significant —including enhanced cooperation, reduced transaction costs, and increased trust in various domains— there are also considerable risks that warrant careful examination. Despite the existence of game-theoretic and economic analyses on the consequences of commitment availability, our understanding of realistic failure scenarios that could come about in the following decades is more shallow.

The present study provides the first comprehensive analysis of such potential negative externalities, or "backfire risks", that may arise from future advancements in commitment technology or their societal implementation. Our main contribution is a categorization of disjunctive "risk stories": qualitatively distinct (yet not mutually exclusive) mechanisms through which advancements in commitment could harm society. While this list is not exhaustive, we do focus on the large-scale catastrophic risks, and discuss those that seem most worrisome and historically plausible.

While not all aspects of this analysis are directly related to AI, it could play a significant role in shaping the future of commitment technologies through multiple routes. Beyond the obvious potential of AI to accelerate the development of more sophisticated implementations of cryptography, specific properties of AI agents or

* Mentoring role. Correspondence to martinsotoq2911@gmail.com

Category	Backfire risk	Example
Risks from rational commitment	Direct threats	Conditionally committing to nuclear warfare
	Informational shutdown	Committing to not read opponent's messages
	Increased volatility	Difficulty in predicting opponent leads to miscoordination
Risks from irrational commitment	Incomplete contracts	Adversarially inserted loophole
	Computational difficulties	Reduced oversight capacity leads to worse regulation
	Disjunctively decentralized commitment ability	Too many government employees have access to the nukes
Technological single point of failure	Deprecation of traditional mechanisms	No centralized diplomacy to fall back on
	Corruptibility of AI-automated mechanisms	Central mechanism is non-reliable or manipulable
Centralization of power		Automated enforcement of precise contracts makes decentralization not required

Our classification of backfire risks from commitment technology, see Section 3.

AI-augmented human institutions could also facilitate more pervasive commitment structures, by enhancing transparency and mediating interactions. The employment of automated negotiation, enforcement and adaptation could transform domains such as international relations, corporate governance, and individual decision-making. The increased role of emerging technologies in contract-making underscores the importance of addressing potential risks.

Given the speculative nature of our subject matter, we take on a macroscopic lens, favoring broad historical and conceptual analysis over detailed case studies or near-term recommendations. The structure of the remainder of the paper is analogous to our sequential methodological approach. In Section 3, we discuss basic definitions and categorizations of commitment mechanisms, and present through historical examples the main drivers of change in their use. In Section 4, we present the backfire risks and discuss their likelihood and severity, through extrapolation of historical trends when possible. In Section 5, we include some final considerations for the further study of these uncertain future risks, as well as tentative avenues for robust governance.

2 Commitment and its history

We will define a commitment as a voluntary punctual action taken by an agent that guarantees (or makes extremely likely, to that agent or others) the enactment of certain additional future actions (by that agent or others). Commitment can be understood as a robust and legible way to ensure certain properties about the future behavior of a certain system (usually, an agent or ensemble of them) through a present action. Commitment technologies are those mechanisms or institutions that enable or improve commitments. We can imagine several different physical **ways in which a commitment can work** (that is, provide a high amount of evidence about those future actions):

1. **By setting strong external incentives** that we assume an agent will comply with. For example, if very powerful legal machinery is making sure that, were Alice to publicly disparage Acme Inc., she

would be forced to pay a huge amount of money (because she signed an NDA previously), then we can safely assume Alice will be disincentivized from doing so.

This showcases that there's not a clear distinction between commitments and more general kinds of incentive-setting (and especially setting the incentives of other agents), and that the strength of the perceived commitment is only proportional to the strength of the incentive.

Another form of setting strong external incentives can be through setting fundamental restrictions on future behavior or possibilities of future behavior, which can be understood as "infinitely strong incentives". For example, Ulysses tying himself to the mast.

2. **By proving internal mechanisms.** For example, if an AI wants to commit to "whenever I see Alice, I will give her 1$", the AI can re-write its own code, and then provide a mathematical proof that, whenever a property X (seeing Alice) holds of its input or internal state, the code's output is "take the action of delivering 1$".

   Less extreme examples include partial transparency into the structure of institutions (which still relies on the incentives of individual teams or members). For example, if you somehow demonstrate that you've spun up inside your company a board tasked with finding corruption cases, and that their members get rewarded for finding more corruption cases, this can be seen externally as a partial commitment to fight corruption.

   Reputation achieves commitment strength through this channel also - there will almost always exist some set of preferences under which the given external disincentives (for example from lower cooperation in future, or legal punishment) would be insufficient to incentivise upholding a commitment (for example if a party had lexicographic preferences for reward in a period in which punishment cannot be applied). As an agent upholds more diverse commitments, the space of preferences under which they could consistently have cooperated then but defect now shrinks - and so given some starting uncertainty about preferences reputation also assists here.

External mechanisms, such as legally binding agreements or informal agreements such as threat of exclusion, provide a formal structure that ensures all parties are held accountable. However, these external measures are often complemented by internal mechanisms embedded within the organizations or systems themselves. For instance, a business partnership may rely on an external contract to outline financial obligations but choose to also rely on internal devices such as reputation or trust to encourage compliance. Such a dual approach might also help strengthen the commitment as by integrating both types, commitments become more robust and capable of addressing multiple dimensions.

At the same time, any of these commitment devices can be **used for different objectives**:

1. **To update the beliefs of other agents.** For example, convince someone you are trustworthy because you have committed to not revealing secrets. Or also, convince someone that if they don't give you something you will attack them (threats).

2. **To lock or incentivize a course of action that would otherwise be hard to guarantee in the future.** For example, Ulysses tying himself to the mast, or using a digital commitment device to incentivize yourself to save money.

Building on these basic definitions, we will be interested in **a more holistic and correlational classification of commitments in a spectrum**, termed the *commitment hierarchy* in past works. The literature reveals a range of classifications, reflecting the varied contexts and purposes these mechanisms serve. A few examples of this categorization include situational vs non-situational commitments (Weinstein 1969), imperative vs motivational commitments (Jan 2003), and soft law vs hard law commitments (Galbraith 2017). Despite this diversity, commitments can broadly be charted inside a spectrum with the following two extremes: low-order and higher-order commitments.

Higher Order Commitments:

Higher order commitments are characterized by their flexibility and responsiveness, allowing for realignment in response to evolving incentives. Often relying on societal mechanisms like trust, reputation, or threat

of exclusion, these commitments are subject to constant reassessment resulting in an unreliable perception (Morrison and Wilhelm Jr 2015) . For example, the New York Stock and Exchange Board used the threat of exclusion to ensure adherence to their rules, motivating compliance (Banner 1998). Their definitory traits are:

1. Flexible and allowing for realignment to changing contexts, incentives, or environments
2. Require reassessment and often reflect of the committing body's immediate interests or preferences
3. Enforced through social contracts, adaptable but also unreliable. Can breed uncertainty, which can be exploited by aggressors for gain

Lower Order Commitments:

Lower order commitments are often characterized by their long-term focus, prioritizing stability and predictability. Enforced with more effort and resources through formal, legal, and structured agreements or contracts, these commitments often acquire a lift of their own and become symbolic of longstanding preferences (Morrison and Wilhelm Jr 2015). For example, the NATO alliance exemplifies such commitment through the mutual defense principle, now signaling reliability and dedication to overarching principles (Sjursen 2004). Their definitory traits are:

1. Stable and reliable, tend to be more rigid and less responsive to evolving contexts
2. Meant to endure over time, transcend immediate circumstances and interests
3. Enforced through sophisticated and legal institutions, rigidity can often lead to fulfillment of irrational and outdated commitments

These categories represent extremes of a spectrum, with most real-world commitments lying somewhere in the middle. Indeed, in many occasions we find commitment procedures with differing properties as related to the definitory properties 1., 2. and 3. above, which don't univocally fall in one end of the spectrum. For example, corporate social responsibility (CSR) initiatives by MNCs are often voluntary and adaptable, allowing companies to adjust based on changing market conditions, reputation concerns, or public relation needs. At the same time, many companies institutionalize CSR as part of their long-term strategies, formalizing commitments to sustainability or ethical practices, and sometimes subjecting them to external audits or regulatory compliance. Thus, CSR combines adaptability with long-term dedication (Sweeney and Coughlan 2013).

And yet, the usefulness of the classification comes from its correlational nature: it is indeed the case that those definitory traits are, most times, found together, due to reinforcing dynamics. For example, in most situations it won't make sense to have a commitment be enforced through sophisticated institutions (3.), while at the same time being very under-specified and requiring of reassessment (1.), since sophisticated institutions (with corresponding high costs) most usually come into play for high-stakes agreements that are important to specify more rigidly.

We'll be especially interested in **the social and technological factors that can change the place on the spectrum of the used commitment methods**. These elements continuously reshape how commitments are enforced, blending flexibility with stability. The discussion below, albeit not exhaustive, attempts to list some of these factors to portray a clearer picture of how commitments advance, and to introduce historical concepts that will prove central to some of our backfire stories in Seciton 3.

Social orderings in a cohesive society are often the basis of higher order commitments such as trust or reputation. This is because such societies rely on informal mechanisms where personal relationships enforce compliance out of aspects like threat of exclusion or social sanctions. However, as societies grow more complex, diverse, and impersonal, these trust-based systems become less effective, necessitating a shift towards more formal, structured systems, such as legal contracts, as society moves down the hierarchy (Morrison and Wilhelm Jr 2015). A historical example of the same would be the industrial revolution. As economies

grew and societal complexity increased, trust-based, reputation-enforced commitments became less feasible. The expanding scale of trade and commerce, and the subsequent migration to big cities where interpersonal relations were unfeasible, required formal contracts and legal mechanisms to manage commitments (Deakin 2005). Similarly, if AI becomes embedded in global governance systems, the complexity and anonymity of interactions could result in a loss of trust, exacerbating challenges in enforcing both legal and informal commitments. This could lead to widespread disillusionment in AI-mediated systems, increasing the risk of backfires, such as public resistance or exploitation of loopholes.

Technological aspects significantly influence commitment mechanisms, either by enhancing or reducing reliance on human intermediaries. Advanced technologies can foster transparency and accountability, promoting higher-order commitments based on trust or community norms. Another impact of improvement in technology is the ability to better record and preserve commitment over longer periods, allowing for greater accountability. The invention of the printing press in the 15th century exemplifies the shift to decentralized information dissemination, allowing for greater transparency and accountability (Dewar 1998). By making it easier to reproduce and distribute documents, the press helped solidify and publicize commitments, moving away from elite-controlled models to more community-based, open systems. The printing press enabled commitments to be recorded and enforced more widely, reducing reliance on elite agreements and fueling movements like the Reformation (Rubin 2014). Online terms and agreements reflect this further. Users consent to terms when they download or use apps, and their consent is stored digitally, preserving the commitment for extended periods without requiring further interaction. These agreements, often extensive and legally binding, are automated and therefore leave no room to negotiate or even fully understand the terms, reducing the flexibility and personal trust involved (Momberg 2016, Limata 2024).

Market information plays a crucial role in determining which type of commitment mechanisms are employed. In situations of high uncertainty, societies often rely on informal mechanisms like trust and reputation to enforce commitments. However, as information becomes more reliable and structured, formal, legally binding commitments become the preferred method for managing relationships and obligations (Morrison and Wilhelm Jr 2015). For example, credit rating agencies like Moody's and Standard & Poor's emerged in the early 20th century to provide more reliable assessment of companies' creditworthiness. Before this, investors often relied on information networks or personal trust to gauge risk, particularly in bond markets (Sylla 2002). The development of rating agencies providing structured and reliable information allowed for more formalized commitments between investors and companies (Allen 1990). As AI systems become more sophisticated in managing and analyzing user data, the reliance on long-term data tracking (like cookies) could increase (I. D. Mitchell 2012). This could lead to more rigid information systems, where users' data and behaviors are tracked indefinitely, with little room for opting out. Just as cookies have raised concerns about privacy and consent, AI-driven data tracking could lead to significant backfires if users feel their personal information is being exploited without sufficient transparency or flexibility.

Legal frameworks provide the foundation for lower-order commitments, which rely on formal, enforceable contracts, and regulations. Strong legal systems ensure that commitments are standardized and respected across borders. However, these legal systems are often slow to adapt to new technological and social challenges, creating gaps in enforcement and accountability. The Nuclear Non-Proliferation Treaty (NPT), signed in 1968, was a legally binding treaty aimed at preventing the spread of nuclear weapons. It relied on formal international legal frameworks and centralized oversight to ensure compliance (Gilligan 2014). While the NPT has been largely successful, non-compliance by states like Iran and North Korea has exposed the limitations of such centralized, formal commitment mechanisms that are rapidly changing geopolitical contexts. Similarly, AI-driven predictive policing, similarly, relies on long-term crime data. As AI systems are becoming more sophisticated in analyzing crime data, many legal systems are starting to formalize their use in predicting criminal activity and allocating law enforcement resources. However, such formalized legal commitments to predictive policing could backfire if the algorithms reinforce biases or make flawed assumptions, leading to over-policing of certain communities or false arrests (Joseph 2024).

3 Backfire risk stories

Although commitment technologies expand the range of possible contracts and bring associated gains, they may also have some backfire risks. They could enable threats, allow ignoring information and increase volatility. They could also interact poorly with incomplete contracts, computational difficulties or the size of institutions. They might also be risky by acting as a single point of failure, weakening the mechanism's effective power and potentially depreciating substitutes.

3.1 Risks from rational commitment

Hawkish strategies, threats and power imbalance. Firstly, stronger commitment technology could enable more threats, both directly by making threats more credible (McClintock, Stech, and Beggan 1987) but also indirectly by making assurances of not carrying out the threatened action more credible. This could act to increase the difference in power between the strongest and weakest states - strong states are currently constrained in large part by weak assurance ability - a "Goliath's curse" - in their ability to extract resources from smaller states (Sechser 2010).

Several other factors may push in the same direction, exacerbating these effects. High productivity countries will disproportionately benefit from AI due to being relatively better endowed with capital than labor, and as high-productivity states are ceteris paribus more powerful, the strongest states will strengthen relative to the weakest (Alonso et al. 2022).[1] If military technology research is done in secret, and if the rate of diffusion does not scale 1-1 with speed-ups in the underlying rate of technological progress, then smaller states will have access to worse military designs than larger ones.[2] In standard models of both democratic and military contexts, power scales as the square of resources employed - so any changes caused by any of these effects could be magnified (Banzhaf III 1964) (Osipov and Helmbold 1991).

Near-term backfire risks from increased crime are also likely. Cybercrime costs peaked in 2002 with 92% of total damages above $800mn to date arising pre-2009 (Johansmeyer 2024). This is in large part due to two commitment difficulties - victims cannot commit to not pursue recovery of funds sent to attackers, and attackers cannot commit to return files or cease attacks upon payment (A. Cartwright and E. Cartwright 2019). Stronger commitment technology could thus enable vastly more damage from cybercrime - until it becomes strong enough for firms to commit to never pay.

Informational shutdown. Secondly, a particular kind of commitment is committing not to take into account certain information - which is dangerous as informational asymmetries are one of the primary drives of conflict, and this could further accelerate commitment races.[3] Warfare is Pareto inefficient over a set of negotiated deals if agents have the same beliefs over war outcome, are risk averse and face no transaction costs. However, if agents have access to different information, both agents could believe that conflict would be better than the bargain they were offered - causing conflict (Fearon 1995).

This is likely to be especially costly if, due to rapid growth induced by AI (or greater short-run variations in military spending in response to higher expected future growth), relative military strength or other capabilities are varying rapidly with time, as this may allow a broader range of possible beliefs about relative capabilities. This could also form an especially damaging part of a commitment race - a situation where multiple agents are racing to implement commitments as soon as possible to choose equilibria advantageous

[1] If population growth fell sufficiently quickly in GDP then this would not be the case: in practice the simple correlation coefficient between GDP/capita and GDP was 0.231, and a Spearman's rank correlation coefficient of 0.483, so this is not a concern.

[2] As this lower diffusion likely implies lower mutual knowledge of capabilities (as without access to a capability the specific technical details could fall in a wide range), this greater informational asymmetry could promote conflict also.

[3] For example, suppose that country A controls an oil field that yields an annual utility of 2, and if the two countries fight over it then B will win - but there is a 50% chance that country B would take a utility cost of 0 and a 50% chance that it would take a utility cost of 8. Country B would rationally take a commitment that resulted in it always not knowing whether the cost would be 0 or 8, verifiably with probability 1-p for p¡0.2, as this would give it expected utility -8p + 2-2p - but expected social utility loss of 4p ¿ 0.

to themselves (Kokotajlo 2019).[4] Even if strong commitment technology was later developed and safely implemented, agents could be permanently limited by the restrictions they had placed on their cognitive processing to make commitments earlier.

Historically, organizations have sometimes been able to avoid internal diffusion of information internally which would be costly in expectation - automotive manufacturers in the 1970s sometimes refused to conduct safety tests as they could only be held liable for a product that had failed to pass those tests (Lyon 2002), and tobacco executives actively suppressed attempts to research the health effects of cigarettes (Oreskes and Conway 2010). This allowed them to make verifiable external commitments that this was the case - such as testifying under penalty of perjury that they were unaware of any evidence that would render it unsafe. However, in these cases organizations mostly pursued these devices due to legal incentives around tort law (Bailey and Baker 2009; K. D. Gilbert and Merrill 2010), rather than directly in order to achieve stronger commitment.

Increased volatility. Thirdly, we know from classical literature in game theory and commitments, as well as recent Computer Science literature in program equilibria (Critch 2022) that introducing contracts or partial transparency to games generally makes the space of equilibria explode. This is obviously possible in allowing for better equilibria, and negative in allowing for worse ones (albeit it can be argued that, at least for now, humans have been successful in mostly taking advantage of the positive side). But there's an additional consideration: the proliferation of equilibria makes choices and behaviors generally harder to predict, and so the whole situation more volatile.

Given human risk-aversion, volatility would seem worse ceteris paribus. Indeed, if agents are risk averse in their monetary payoffs, and the average change across equilibria for an agent's payoffs is zero or negative, then this will lower overall welfare in expectation - and lowering the returns to investment could lower the payoffs in all worlds, even those where the equilibrium doesn't change.

Research is nascent and ongoing on the exact consequences of these capabilities for computationally bounded agents. We are especially uncertain about the economic interaction of agents with different computational bounds when such an array of contracts becomes available. We could worry that the more intelligent agent will always be able to find methods or proposals to exploit the rest. While there is hope that with transparency we can construct pragmatic gradients towards the newly opened best equilibria, the rapid global availability of fine-grained commitments or transparency is a game-theoretic state of affairs which we understand even less, and whose possibly high-variance evolution we cannot predict as of yet.

3.2 Risks from irrational commitment

Incomplete contracts. Many relations are today managed by incomplete contracts, with reputation bridging the gap. If agents are inexperienced in translating their incomplete to complete contracts, then significant and persistent costs could result.

Incomplete specifications are a general problem in the design of safe systems, not only related to commitments. Attempts to generate provably safe AI (see Dalrymple et al. 2024) rely on the specification of a safety function; simple reinforcement learning relies on explicitly specifying a human reward function. ARIA is currently spending £59mn on providing a detailed world specification (Advanced Research and Invention Agency (ARIA) 2023). As we have previously failed to correctly specify incentives in even extremely specific contexts, creating a fully general specification is unlikely to be achieved correctly the first time - but competitive pressures may ensure that such AI systems are given substantial power even if imperfect (Pan, Bhatia, and Steinhardt 2022). As more precise specifications may also misgeneralise when vaguer ones would not, greater volatility may exacerbate this.

The canonical example of transition would be increasing the specificity of incentives within contracts -

[4]For example, in the game chicken it is in both players interest to be the first to verifiably lock themselves into never swerving.

specifically rewarding hospitals for having lower waiting times, nail factories for more nails, or viper-catchers for bringing more dead vipers back. However, Goodhart's Law poses a substantial barrier here - "when a measure becomes a target, it ceases to be a good measure" (Strathern 1997). In practice, hospitals had ambulances idle outside to avoid starting the clock (Bevan and Hood 2006), nail factories produced unusably small nails and viper catchers began breeding the snakes themselves - and when the scheme was terminated, released the now valueless vertebrates en masse (H. Mitchell 2004). Historically, such transitions were often simultaneous - by 2002 Tony Blair's labor administration had over 300 targets across government departments (Hood 2006) - suggesting a threat channel from inexperienced agents updating a larger number of such contracts to be more specific simultaneously.

Computational difficulties. For some decisions, forecasting their impact on the world in advance is hard - and so decision makers could simply get it wrong. This is again more likely in higher-volatility worlds - and so more likely in the future. If AI substantially increases the speed of warfare, then this could also reduce any human decision makers' time for contemplation greatly - although many such periods are already short.[5] Even in the speculative presence of advanced technology such as whole brain emulation, faster AI contemplation is likely because uploads of human brains could likely be sped up at only a linear increase in run-cost (Hanson 2016) - and this allows reductions in latency not just increases in throughput so is likely to be utilised. If non-uploaded human control is continued, this may thus lead to reduced contemplative time relative to underlying processes.[6]

Institutional size. Distributed commitment ability could result in institutions substantially overcommitting. The stronger the commitment technology that can be accessed by any individual, the greater the degree of overcommitment.

AI institutions may be much larger - wages today exceed subsistence many times over, so for many priceability combinations total population of an agent with a low fixed cost relative to marginal cost of instantiation would increase. [7] This must lead to some combination of an increase in the number of agents per organization or in the number of organizations. When either occurs, firms become more decentralized (Bloom, Sadun, and Van Reenen 2010) - resulting in diffusion of commitment power amongst the org, with the top functional management team also growing larger (Guadalupe, H. Li, and Wulf 2014) having the same effect. [8]

Competition will also likely increase in the future due to more elastic supplies of factors of production, especially labor (Hanson 2016). Additionally, such changes may promote conflict as individuals within an organization are often promoted for choosing actions which benefit the organization, but organizations often lack punishment options more extreme than firing the individual. This generalises to all cases where a principal will inflict maximum punishment on an agent for all damage amounts beyond some threshold, and is analogous to liability problems in AI risk (see Trout 2024). This means that individuals have incentives to be hawkish internally as such policies are often higher variance than the (known) cooperative payoff. [9] [10]

[5] In the event of a nuclear strike on the US, the President would on paper have 12 minutes to make a retaliation decision before many assets were destroyed - submarines in port, bombers on the ground and missiles in land-based siloes. At any given time this accounts for most American deployed warheads. However, in practice they may have much less, given that the Secret Service may wish to use much of that window to get them to as much safety as possible (Jacobsen 2024).

[6] If uploaded humans control such worlds, then the reverse could be true - the subjective time between a nuclear launch and receiving a strike could be several lifetimes (Hanson 2016).

[7] An analogous argument is that if individual AI agents remain packaged at roughly human productivity per instance - as seems to be the case with current language models - then institutional size will greatly increase as the economy expands. This is because existing AI mostly offers not productivity improvements to manufacturing, but instead a new form of capital that substitutes with service provision. Production can only rise as rapidly as it does in Hanson 2001 - doubling every 18 months - because the agent population is rising at a similar rate.

[8] For qualitative descriptions of the same phenomena see Williamson 1975, March 1994 and Bebchuk and Roe 1999b

[9] If two agents are using an agreed fair bargaining algorithm (e.g. Nash et al. 1950), do so on the basis of their common knowledge about each others capabilities and cannot add reveal new information following a decision to resolve a conflict cooperatively, then the variance of the cooperative payoff is zero while war will have positive variance due to both sides maintaining private capabilities information (and likely also due to inherent randomness in conflict (C. v. Clausewitz 1976)). Toy examples can however be constructed against this - defecting will often be lower variance than cooperating on the prisoners dilemma for some beliefs about an opponent's choices (for example in the defect-defect equilibrium).

[10] S. Li and Schmitz 2020; Kroumi, Lessard, and Soares 2021; Kroumi and Lessard 2021 all study the effects of differences in variance between payoffs, and find that cooperation is more attractive if it is relatively lower variance than defection, ceteris

Aggregate commitment power will thus rise, regardless of changes in strength in the underlying commitment technology, exacerbating any negative effects.

More decentralized institutions will also exercise their stronger commitment power because individuals have heterogeneous preferences - and empirically do not employ conditional reasoning. Rational agents should require stronger evidence to impose a unilateral policy if more agents have the ability to impose the policy; empirically humans do not, relying on the same evidence threshold in a simple game with 2 or 20 players (Lewis et al. 2024). Additionally, if agents have differing preferences then the more agents have commitment power the more likely a commitment option will satisfy at least one such set of preferences.

3.3 Single points of failure

The technology could pose a single point of failure, due to potentially centralising the execution of commitments into a single system, and be vulnerable to simple technical failure or a malicious actor. The weakness of single points of failure - such as dictators - is today masked by the large blowback in the event of failed assassination attempts. Jones and Olken 2009. If attacks against software become less visible than assassination attempts - because attackers can search for vulnerabilities without defenders knowing, organizations rarely report attack attempts and attribution is very challenging - this blowback barrier would be removed.

Corruptibility of AI-automated mechanisms. A single system could be vulnerable in the event of conflict, threatening to turn any small conflict into a global one if a large number of actors had the capability to disrupt the mechanism. If most commitments within a country were resolved through a single piece of software, other states would be able to inflict substantial economic disruption by attacking such software, and so would likely work hard to learn several zero-day vulnerabilities. Given that power volatility is thought of as one of the primary causes of conflict (Sample 2018; Hebron, James, and Rudy 2007; Geller 1992), states may be especially likely to leverage this to cause damage.[11] In the limit, the commitment technology would not function as such if multiple actors had the ability to interrupt its operation if it resolved unfavorably to them - limiting the kinds of contracts it could perform.

This may be analogous to cartels - every firm faces very large incentives to cheat and defect as optimal production holding constant that of others is much higher. OPEC has survived from the mid-20th century, but not the International Coffee Agreement, International Tin Council, International Natural Rubber Agreement or the International Copper Cartel (see Akiyama and Varangis 1990; Cheyne 1989; Tharian George 1987; Guzmán 2018). Other examples of similarly unstable structures might be global currency systems where members can freely devalue, such as the 1920s Gold Standard (Obstfeld and Taylor 2003).

Deprecation of other mechanisms like diplomacy or reputation. Diplomatic technology had to be invented, with the classical period featuring state breakout much more frequently than the early modern period as a consequence.[12] If more disputes were handled through stronger commitment technology, then traditional diplomatic technology could be depreciated, worsening the resolution of any processes outside the technology's reach - or, in the event of the technology's failure, worsening the outcomes relative to no depreciation.

Democracies often have sharp breaks in foreign policies between administrations, so institutional knowledge amongst senior officials can decay quickly (Hunt 2009; Sarotte 2014). Additionally, while it might be straightforward for diplomacy to adapt gradually to changes in the world, it may be unable to do so if it has to make a larger adjustment simultaneously - which means that even a short break in continuing traditional diplomacy in a volatile world could have persist effects on the quality of negotiation available. Lastly, if these commitment mechanisms are especially likely to fail during periods of heightened tensions - for example if they are a single point of failure compromised by enemy action - then risks may also be higher.

paribus - but they don't explicitly discuss the variance of cooperation versus defection in general.

[11] Of course, with completely perfect commitment technology power volatility ceases to cause conflict (Debs 2024).

[12] For example, states were often extremely explicit about the status of subject communities, rendering diplomacy difficult - Athenians referred to their Delian League subjects as "those poleis which the Athenians rule". Low 2005

4 Recommendations for policy and further research

Given our highly uncertain state of knowledge, the first and foremost recommendation remains deepening the study of these failure modes to more thoroughly understand which dynamics seem fundamentally dangerous, and which interventions robustly positive. That said, we present here some tentative directions for future interventions that as of now seem most promising.

Commitment technology should be transparent. Races associated with stronger commitment technology have high downside risks when one side makes commitments without the knowledge of the other. Greater transparency of the level of commitment technology mitigates this - if one agent commits to never swerve in chicken, then the other agent will not commit the same if they are aware of this. However, this does also increase the incentive to develop such technology ex ante, as the main downside risk from doing so is removed - increasing the other backfire risks.

Organizations should explicitly track their commitment ability. As individuals do not appear to employ conditional reasoning and have heterogeneous preferences, organizations can currently achieve levels of commitment ability much stronger than would be available by any single individual having power by giving the same power to more individuals. If the strength of commitment technology available increases, the optimal number of individuals to wield it decreases - and so organizations should explicitly track their desired level to ensure that they reduce the number of individuals with access to stronger forms of the technology.

Mandatory insurance for critical commitment technology. If a centralised commitment mechanism run by a limited liability firm failed, it might cause damages much greater than the firm's valuation. This would resulting in the firm facing inadequate incentives to ensure safety. Mandating insurance should ensure that much larger damage amounts would be covered via insurers adjusting premia in response to risk levels, rendering safety incentives closer to socially optimal (Shavell 2007). Tracking failed attempts, for example via canary tokens, would also be crucial to minimise the risk of overturning the current channel by which assasination attempts have little impact in expectation.

Employing commitment technology itself for technological governance. As commitment technologies advance, they could be leveraged to create more robust governance mechanisms for potentially risky technologies, including AI and widely-available commitment technology itself. These could include verifiable development milestones, automatic restrictions based on safety conditions, enforced transparency agreements, and global coordination mechanisms. Indeed, transparency and binding agreements have in the past helped restrict the downsides from rapid technological adoption. However, implementation should be gradual and carefully monitored, as the same risks identified in this paper could apply to these governance mechanisms as well.

5 Conclusion

We have provided a first comprehensive analysis of potential "backfire risks" arising from future advancements in commitment technology. While stronger commitment mechanisms offer significant benefits in enhanced cooperation and reduced transaction costs, they also present considerable risks. These include the exacerbation of well-known dangers with historical precedent, as well more radically new technical possibilities. The rapid advancement of AI is likely to play a significant role in shaping these technologies, potentially accelerating their viability, development and implementation. Given the speculative nature of our subject matter, we emphasize the critical need for further research in areas such as methods for organizations to manage their commitment abilities, the interplay between AI advancements and commitment technologies, or the purposing of the technologies themselves for regulation. As these technologies continue to evolve, it is crucial that researchers, technologists, and policymakers collaborate to develop robust governance frameworks and safety measures, balancing the potential benefits with the mitigation of risks.

References

Advanced Research and Invention Agency (ARIA) (2023). *Safeguarded AI Programme*. https://www.aria.org.uk/programme-safeguarded-ai/.

Akiyama, Takamasa and Panayotis N Varangis (1990). "The impact of the International Coffee Agreement on producing countries". In: *The World Bank Economic Review* 4.2, pp. 157–173.

Allen, Franklin (1990). "The market for information and the origin of financial intermediation". In: *Journal of financial intermediation* 1.1, pp. 3–30.

Alonso, Cristian et al. (2022). "Will the AI revolution cause a great divergence?" In: *Journal of monetary economics* 127, pp. 18–37.

Amodei, Dario et al. (2016). "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565*.

Anderson, Ross et al. (2013). "Measuring the cost of cybercrime". In: *The economics of information security and privacy*, pp. 265–300.

Ashley Leeds, Brett (2000). "Credible commitments and international cooperation: Guaranteeing contracts without external enforcement". In: *Conflict Management and Peace Science* 18.1, pp. 49–71.

Bahrami, Bahador et al. (2010). "Optimally interacting minds". In: *Science* 329.5995, pp. 1081–1085.

Bailey, Albert F. and Jeremy J. Baker (2009). "Corporate Misconduct and Tort Law". In: *Journal of Legal Studies* 38.2, pp. 425–451. DOI: 10.1086/598178.

Banner, Stuart (1998). "The origin of the New York stock exchange, 1791–1860". In: *The Journal of Legal Studies* 27.1, pp. 113–140.

Banzhaf III, John F (1964). "Weighted voting doesn't work: A mathematical analysis". In: *Rutgers L. Rev.* 19, p. 317.

Bebchuk, Lucian Arye and Mark J Roe (1999a). "A theory of path dependence in corporate ownership and governance". In: *Stanford Law Review*, pp. 127–170.

— (1999b). "A theory of path dependence in corporate ownership and governance". In: *Stan. L. Rev.* 52, p. 127.

Bevan, Gwyn and Christopher Hood (2006). "Performance measurement in the NHS: much ado about nothing?" In: *Public Money & Management* 26.2, pp. 117–120.

Blair, Bruce G (2011). "Global Zero Alert for Nuclear Forces". In: *The Brookings Institution*.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen (2010). "Recent advances in the empirics of organizational economics". In: *Annu. Rev. Econ.* 2.1, pp. 105–137.

Cartwright, Anna and Edward Cartwright (2019). "Ransomware and reputation". In: *Games* 10.2, p. 26.

Cheyne, Ilona (1989). "The International Tin Council". In: *The International and Comparative Law Quarterly* 38.2, pp. 417–424.

Clausewitz, Carl von (1976). *On War*. English. Trans. by Michael Howard and Peter Paret. Princeton, NJ: Princeton University Press. ISBN: 978-0691018856.

— (1989). *On War*. Princeton University Press.

Critch, Andrew (2022). "Surprises and problems in open-source game theory". In.

Dalrymple, David et al. (2024). "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems". In: *arXiv preprint arXiv:2405.06624*.

Deakin, SF (2005). *The Law of the Labour Market: Industrialization, Employment, and Legal Evolution*.

Debs, Alexandre (2024). "Assurance and Self-Assurance under Power Imbalance". In.

Dewar, James A (1998). *The information age and the printing press: Looking backward to see ahead*.

Eichengreen, Barry (1992). *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939*. Oxford University Press.

Fearon, James D (1995). "Rationalist explanations for war". In: *International Organization* 49.3, pp. 379–414.

Galbraith, Jean (2017). "From Treaties to International Commitments: The Changing Landscape of Foreign Relations Law". In: *U. chI. l. rev.* 84, p. 1675.

Geller, Daniel S (1992). "Power transition and conflict initiation". In: *Conflict Management and Peace Science* 12.1, pp. 1–16.

Gilbert, Christopher L (1987). "International commodity control: Retrospect and prospect". In: *The World Bank Research Observer* 2.2, pp. 243–263.

Gilbert, Kevin D. and Max H. Merrill (2010). *Criminal Law and the Corporate Officer*. New York, NY: Aspen Publishers.
Gilligan, Kimberly (2014). "The non-proliferation regime and the NPT". In: *Nuclear Non-Proliferation in International Law-Volume I*, pp. 85–104.
Guadalupe, Maria, Hongyi Li, and Julie Wulf (2014). "Who lives in the C-suite? Organizational structure and the division of labor in top management". In: *Management Science* 60.4, pp. 824–844.
Guzmán, Juan Ignacio (2018). "The International Copper Cartel, 1935–1939: the good cartel?" In: *Mineral Economics* 31, pp. 113–125.
Hanson, Robin (2001). *Economic growth given machine intelligence*. Tech. rep. Citeseer.
— (2016). *The age of Em: Work, love, and life when robots rule the earth*. Oxford University Press.
Hebron, Lui, Patrick James, and Michael Rudy (2007). "Testing dynamic theories of conflict: Power cycles, power transitions, foreign policy crises and militarized interstate disputes". In: *International Interactions* 33.1, pp. 1–29.
Hood, Christopher (2006). "Gaming in targetworld: The targets approach to managing British public services". In: *Public Administration Review* 66.4, pp. 515–521.
Houweling, Henk and Jan G Siccama (1988). "Power transitions as a cause of war". In: *Journal of Conflict Resolution* 32.1, pp. 87–102.
Hunt, Michael H (2009). *Ideology and US foreign policy*. Yale University Press.
Jacobsen, Annie (2024). *Nuclear War: A Scenario*. Penguin.
Jan, Stephen (2003). "A perspective on the analysis of credible commitment and myopia in health sector decision making". In: *Health Policy* 63.3, pp. 269–278.
Johansmeyer, Tom (2024). "Surprising stats: the wrost economic losses from cyber catastrophes". In: *The Loop: ECPR*. URL: https://theloop.ecpr.eu/surprising-stats-the-worst-economic-losses-from-cyber-catastrophes/.
Jones, Benjamin F and Benjamin A Olken (2009). "Hit or miss? The effect of assassinations on institutions and war". In: *American Economic Journal: Macroeconomics* 1.2, pp. 55–87.
Joseph, Jeena (2024). "Predicting crime or perpetuating bias? The AI dilemma". In: *AI & SOCIETY*, pp. 1–3.
Kimbrough, Erik O et al. (2015). "Commitment problems in conflict resolution". In: *Journal of Economic Behavior & Organization* 112, pp. 33–45.
Kokotajlo, Daniel (2019). *The "Commitment Races" Problem*. URL: https://www.lesswrong.com/posts/brXr7PJ2W4Na2EW2q/the-commitment-races-problem.
Kroumi, Djamel and Sabin Lessard (2021). "Evolution of cooperation under payoff variability". In: *Dynamic Games and Applications* 11.3, pp. 707–726.
Kroumi, Djamel, Sabin Lessard, and Carlos Soares (2021). "Variance in payoffs and the evolution of cooperation in finite populations". In: *Journal of Theoretical Biology* 512, p. 110561.
Lewis, Joshua et al. (2024). "It Only Takes One: The Psychology of Unilateral Decisions". In.
Li, Shengwu and Patrick W Schmitz (2020). "Cooperation and selection under risk and ambiguity". In: *Journal of Economic Behavior & Organization* 180, pp. 1–15.
Limata, Plinio (2024). "Blockchain and institutions: trust and (de) centralization". In: *International Review of Economics* 71.1, pp. 1–17.
Low, Polly (2005). "Looking for the language of Athenian Imperialism". In: *The Journal of Hellenic Studies* 125, pp. 93–111.
Lyon, Thomas P. (2002). "Law and Public Health Regulation of the Automobile". In: *Annual Review of Public Health* 23, pp. 349–369. DOI: 10.1146/annurev.publhealth.23.100901.140437.
March, James G (1994). *A primer on decision making: How decisions happen*. Simon and Schuster.
— (1999). *The Pursuit of Organizational Intelligence*. Blackwell Publishers.
McClintock, Charles G, Frank J Stech, and James K Beggan (1987). "The effects of commitment to threats and promises upon bargaining behaviour and outcomes". In: *European Journal of Social Psychology* 17.4, pp. 447–464.
Mitchell, Horace (2004). *The Cobra Effect: Stories of Unintended Consequences*. Cato Institute.
Mitchell, Ian D (2012). "Third-party tracking cookies and data privacy". In.
Momberg, Rodrigo (2016). "Standard terms and transparency in online contracts". In: *Standard Terms and Transparency in Online Contracts. Intersentia*, pp. 189–206.

Morrison, Alan D and William J Wilhelm Jr (2015). "Trust, reputation, and law: the evolution of commitment in investment banking". In: *Journal of Legal Analysis* 7.2, pp. 363–420.
Nash, John F et al. (1950). "The bargaining problem". In: *Econometrica* 18.2, pp. 155–162.
Obstfeld, Maurice and Alan M Taylor (2003). "Sovereign risk, credibility and the gold standard: 1870–1913 versus 1925–31". In: *The Economic Journal* 113.487, pp. 241–275.
Oreskes, Naomi and Erik M. Conway (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York, NY: Bloomsbury Press.
Osipov, M and Robert L Helmbold (1991). *The influence of the numerical strength of engaged forces on their casualties*. US Army Concepts Analysis Agency.
Pan, Alexander, Kush Bhatia, and Jacob Steinhardt (2022). "The effects of reward misspecification: Mapping and mitigating misaligned models". In: *arXiv preprint arXiv:2201.03544*.
Perrow, Charles (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books.
Powell, Robert (2006). "War as a commitment problem". In: *International organization* 60.1, pp. 169–203.
Proctor, Robert N (2012). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*. University of California Press.
Rubin, Jared (2014). "Printing and Protestants: an empirical test of the role of printing in the Reformation". In: *Review of Economics and Statistics* 96.2, pp. 270–286.
Sample, Susan G (2018). "Power, wealth, and satisfaction: When do power transitions lead to conflict?" In: *Journal of Conflict Resolution* 62.9, pp. 1905–1931.
Sarotte, Mary Elise (2014). "A broken promise? What the West really told Moscow about NATO expansion". In: *Foreign Affairs* 93, pp. 90–97.
Schelling, Thomas C (1960). *The Strategy of Conflict*. Harvard University Press.
Sechser, Todd S (2010). "Goliath's curse: Coercive threats and asymmetric power". In: *International Organization* 64.4, pp. 627–660.
Shavell, Steven (2007). "Liability for accidents". In: *Handbook of law and economics* 1, pp. 139–182.
Sjursen, Helene (2004). "On the Identity of NATO". In: *International Affairs* 80.4, pp. 687–703.
Strathern, Marilyn (1997). "'Improving ratings': audit in the British university system". In: *European Review* 5.3, pp. 305–321.
Sun, Xinyuan et al. (2023). "Cooperative AI via Decentralized Commitment Devices". In: *arXiv preprint arXiv:2311.07815*.
Sweeney, Lorraine and Joseph P Coughlan (2013). "Corporate Social Responsibility and Firm Performance: A Stakeholder Approach". In: *Academy of Management Proceedings*. Vol. 2013. 1. Academy of Management Briarcliff Manor, NY 10510, p. 14466.
Sylla, Richard (2002). "An historical primer on the business of credit rating". In: *Ratings, rating agencies and the global financial system*, pp. 19–40.
Tharian George, K (1987). "International Commodity Agreements: The Case of Natural Rubber". In: *Social Scientist*, pp. 77–86.
Trout, Cristian (2024). "Liability and Insurance for Catastrophic Losses: the Nuclear Power Precedent and Lessons for AI". In: *arXiv preprint arXiv:2409.06673*.
Victor, David G, Marcel Lumkowsky, and Astrid Dannenberg (2022). "Determining the credibility of commitments in international climate policy". In: *Nature Climate Change* 12.9, pp. 793–800.
Weinstein, Franklin B (1969). "The concept of a commitment in international relations". In: *Journal of Conflict Resolution* 13.1, pp. 39–56.
Williamson, Oliver E (1975). "Markets and hierarchies: analysis and antitrust implications: a study in the economics of internal organization". In: *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.