Secret Innovation

June 7, 2024

Abstract

Conventional wisdom holds that open, collaborative, and transparent organizations are innovative. But some of the most radical innovations—satellites, lithium-ion batteries, the Internet—were conceived by small, secretive teams in national security agencies. Are these organizations more innovative because of their secrecy or in spite of it? We study a principal-agent model of public-sector innovation. We give research teams a secret and public option during the initial testing and prototyping phase. Secrecy helps advance high-risk, high-reward projects through the early phase via a cost-passing mechanism. In open institutions, managers will not approve pilot research into high-risk, high-reward ideas for fear of incurring political costs. Researchers exploit secrecy to conduct pilot research at a higher personal cost to generate evidence that their project is viable and win their manager's approval. Contrary to standard principal-agent findings, we show that researchers may exploit secrecy even if their preferences are perfectly aligned with their manager's; and that managers do not monitor researchers even if monitoring is costless and perfect. We illustrate our theory on two cases from the early Cold War: the CIA's attempt to master mind control (MKULTRA) and the origins of the reconnaissance satellite (CORONA). We contribute to the political application of principal-agent theory and studies of national security innovation, emerging technologies, democratic oversight, and great power competition.

1 Introduction

Nations that want prosperity and security must innovate (Horowitz and Pindyck, 2023; Taylor, 2016). Scholars across many disciplines study why some organizations are innovative. There is broad agreement that openness, defined loosely as organizations that encourage employees to internally share ideas with colleagues, spurs innovation (Laursen and Foss, 2003; West and Anderson, 1996). Internally open organizations foster competition and collaboration between otherwise siloed divisions (Aghion, Bloom, Blundell, Griffith and Howitt, 2005; Macdonald, 2015), diffuse ideas (Boushey, 2016), and encourage a free-flow of information (Zoghi, Mohr and Meyer, 2010).

One set of institutions consistently buck this trend: secretive intelligence and national security organizations. Agencies like the Central Intelligence Agency (CIA), the Defense Advanced Research Projects Agency (DARPA), and MI6 often discourage internal sharing, but have consistently produced radical innovations. These include the satellite (Oder, Fitzpatrick and Worthman, 1988), autonomous robots (Jacobsen, 2015), and lithium-iodine batteries (Richelson, 2002, 257). Failed projects—from turning cats into listening devices to psychic spying—also speak to their vision (Houghton, 2019; Richelson, 2002).

Are these organizations more innovative because of secrecy or in spite of it? To answer this question, we study a principal-agent model of organizational innovation (e.g Lai, Riezman and Wang, 2009; Kopel and Riegler, 2006). We adapt the payoffs to reflect the actors' sensitivities to political costs and benefits (Joseph, Poznansky and Spaniel, 2022). We allow researchers secrecy during the conceptualization of innovation. We then contrast mechanisms for innovation in open and secret public-sector institutions (Cain, 2014).

Secrecy allows actors to distribute the political costs of authorizing each phase of politically sensitive research. In open institutions, lower-level researchers cannot pursue pilot programs to determine if a concept is viable without their manager learning about it. When a manager learns of a novel but controversial idea, she will not even approve pilot research because she does not want to be responsible. Secrecy early in the innovation process gives an enterprising researcher cover to collect evidence (at a larger personal cost) that the novel idea is viable. If it shows promise, the researcher seeks manager approval.

Our mechanism generates two surprising results for international relations principal-agent the-

ory (Downs and Rocke, 1994; Hawkins, Lake, Nielson and Tierney, 2006; Lonardo, Sun and Tyson, 2020; Malis, 2021, 2024). First, the researcher turns to secrecy even if her preferences are perfectly aligned with the manager's. Second, the manager does not monitor the researcher even if monitoring is costless and perfect—*and* the manager knows that the researcher only exploits secrecy to do something the manager would not allow her to do. These results follow from a don't-ask-don't-tell dynamic made possible because secrecy allows actors to distribute costs. Distributing costs alleviates preference asymmetry. The manager knows if she monitors the researcher, she will discover something unsavory and shut research down. However, if she remains ignorant, she can incur a small share of the costs associated with highly controversial pilot research and still benefit when the pilot research shows promise.

We find secretive national security institutions should produce different kinds of innovations on average. All political organizations pursue ideas that in expectation serve the national interest and involve non-controversial research practices. However, only secret organizations can pursue *initial concepts* that involve large risks and rewards. Before pilot research is carried out, these ideas are too controversial for open, public-sector organizations to pursue. With hindsight, they represent both the path-breaking innovations the intelligence community is known for, and some of its shameful failures.

We illustrate our theory using several cases: attempts to master mind control (MKULTRA) and innovations to facilitate reconnaissance (via the CORONA satellite and, in the qualitative appendix, the U-2 spy plane). We chose these cases because of historical features that provide inferential leverage. But they are also important for, but overlooked by, international relations scholars. MKULTRA's exposure in the 1970s affected intelligence reforms and legislative oversight for decades afterwards. The case is especially important given that, according to secondary accounts, the CIA and DARPA recently looked into brainwashing (Jacobsen, 2015, 109). Existing studies examine the effects of reconnaissance technologies on conflict dynamics (Carnegie and Carson, 2018; Coe and Vaynman, 2019; Early and Gartzke, 2021; Vaynman, 2022), but overlook what it took to develop them. We show these research programs may have been shelved but for internal secrecy.

We contribute to several debates. First, we provide a logic for the national security origins of radical innovations. For international security scholars, this clarifies selection into technology shocks that can cause (eg Debs and Monteiro, 2014) and offset (Coe and Vaynman, 2019) conflict.

It also helps explain the conditions under which state-led, highly productive innovations of interest to IPE scholars occur (e.g, Drezner, 2019; Farrell and Newman, 2019; Miller, 2022; Taylor, 2016).[1]

Second, we refine and integrate theories of military innovation (Kuo, 2022; Posen, 1984; Rosen, 1988). Conventional wisdom is that while "militaries have strong incentives to innovate to succeed in war," they are "slow to innovate" because of hierarchical structures and entrenched interests (Neads, Farrell and Galbreath, 2023). We connect agency-wide incentives to respond to international threats with individual-level incentives for innovation (Jungdahl and Macdonald, 2015). We highlight an unexplored bureaucratic feature: agencies that practice internal secrecy owing to fear of foreign rivals. Since the level of internal secrecy varies across organizations and time, we also explain unexplored variation in peacetime innovation. Moreover, we broaden this research to agencies beyond the military, and focus more on technological rather than doctrinal innovation.

Third, we contribute to secrecy research in international relations. Most find that secrecy reduces welfare (see Carnegie, 2021) because it creates uncertainty at the international level (e.g. Wolford, Reiter and Carrubba, 2011; Joseph and Poznansky, 2018; Carnegie and Carson, 2018), and facilitates domestic inefficiencies (Colaresi, 2012; Goldfien and Joseph, 2023; Goldfien, Joseph and Krcmaric, 2023). We expand the limited research on efficient secrecy (Carson, 2018; Kurizaki, 2007; Poznansky, 2020; Joseph, 2021) by showing that it allows states to pursue welfare-enhancing projects, not only avoid costs. We also define internal and external secrecy and explain their connection.

2 Concepts

Our theory is closest to principal-agent models that examine rationalist, organizational innovation (e.g Lai et al., 2009; Kopel and Riegler, 2006). We adapt this model to fit public-sector employees, and secrecy. Others examine principal-agent problems in political institutions (see Miller, 2005). But they typically focus on policymaking or electoral accountability and not innovation (Downs and Rocke, 1994, is closest to this study). Some examine principal-agent problems unique to international relations. But they emphasize interactions between two states (Hawkins et al., 2006) or foreign militaries (Biddle, Macdonald and Baker, 2018). We focus on a handful

[1] Some estimate that satellites have contributed over a trillion dollars to the US economy. Micro-chips, lithium-ion batteries, and the green revolution hold similar implications.

of employees working within government agencies. We detail the differences in Appendix D. Our theory shares a substantive focus with national security innovation. But our theoretical approach is different. We review how we complement this literature in Appendix C. Here we further develop our two central concepts: innovation and secrecy.

2.1 Innovation

Innovation is the process of taking a novel idea and converting into a working device or policy (Kollars, 2017; Taylor, 2016, 126). Innovation occurs only after (1) a novel idea, (2) pilot testing to validate and improve that insight, and (3) the decision to develop a product and deploy it in the field (King, 1990).[2] The last step is critical. It is not enough to conceive an idea. Innovation requires that it is developed into a working product (West and Anderson, 1996).

Government agencies innovate to achieve their policy goals (Taylor, 2016). Consistent with others, we define an innovation's effects as whether the final product advances the nation's goals such as improved military effectiveness, intelligence collection, and enhanced security and prosperity (Horowitz and Pindyck, 2023).[3] Many innovations have positive effects (i.e. move the organization towards its goals). Others have no effect. Others still have negative effects because of unintended consequences (Sechser, Narang and Talmadge, 2019; Joseph, 2023). This could include conflict escalation, degrading defenses, or facilitating local rebellion (Horowitz, 2020; Kuo, 2020). While researchers hold expectations, they are uncertain about the true effect.

We distinguish between the effects of innovation and the costs of moving an idea through development phases (King, 1990). Some development costs stem from the financial burden of research trials and prototype construction. But public-sector institutions are especially sensitive to political costs.[4] These can manifest at different stages of innovation. During pilot research, political costs can come from wasteful spending or human subjects research without consent. During the deployment phase, they can follow from labor abuses during production or escalation with rivals.

Of course, not all research activates political costs.[5] But in many cases, national security

[2] See Horowitz and Pindyck (2023).

[3] We bracket distributional concerns because our actors are national security professionals who typically hold higher public service motivations than the average citizen (see Houston, 2000).

[4] Private and public organizations both accrue political and financial costs. However, firms mainly consider the financial liabilities of both costs and effects (see King, 1990).

[5] Our model accounts for this because we allow costs to be 0.

employees do face costs, including from organizational cultures that perceive radical ideas as reckless (Grissom, 2006; Lee, 2019). Since they are dealing with public funds, risky spending (which is promoted in private firms) is often viewed as a violation of federal code of conduct under the waste, fraud and abuse standard. Unique ethical concerns surrounding the use of force impose distinct personal costs on national security innovators (Zhang, Anderljung, Kahn, Dreksler, Horowitz and Dafoe, 2021). Many ideas also raise the risk of tragic accidents wherein soldiers die, or test satellites crash into foreign territory. When this happens, investigators scrutinize those who plan and approve these programs looking for mistakes. These anticipated costs can be large enough that researchers do not voice their ideas in the first place. This helps explain why militaries may fail to pursue novel ideas even though the problems are important and their budgets are large.

Later, these contextualizing details will help us interpret our theoretical findings. But in the end, our model is abstract. We only assume that different public-sector employees participate in the research process and derive benefits (positive or negative) depending on the effects of innovation. They also incur research and development costs as ideas go through the innovation process. The scope of these costs depends on how responsible they are for advancing an idea and their personal sensitivities.

2.2 Internal Secrecy

While democracies promote scrutiny of government agencies, national security agencies enjoy a special status. Specifically, they are allowed to keep secrets owing to fears of foreign threats (Colaresi, 2014). We refer to this phenomenon as external secrecy. To sustain external secrecy, national security agencies often practice internal secrecy. That is, certain individuals and groups do not need to, or are discouraged from, sharing information with colleagues, those who hold oversight over them, or even their own superiors.[6]

Internal secrecy is necessary to sustain external secrecy for two reasons. First, foreign threats infiltrate national security agencies. Foreign penetration can be stemmed by limiting who knows important facts. Second, whistleblowers and leakers have historically revealed large document corpuses without fully understanding the potential for national security harm. By restricting access,

[6]Delegation is an important part of secrecy in our theory. Others have studied delegation in private sector innovation, but find it only enhances innovation because information is liberally shared (Jones, Kalmi and Kauhanen, 2006; West and Anderson, 1996; Aghion et al., 2005).

agencies limit what they release publicly (CIA, 1960).

Some aspects of internal secrecy are institutionalized. In the US, for example, program officers must alert contract officers to purchases, who then openly tender contracts. But "[f]ull and open competition need not be provided for when the disclosure of the agency's needs would compromise the national security."[7] Other aspects of internal secrecy follow from a culture of need-to-know. Because of the "sensitive nature of their work, intelligence organizations have been reluctant to engage in bidirectional dialogue with decision-makers and the larger public" (Ivan, Chiru and Arcos, 2021, 505).

Different still, monitoring and evaluation is mandatory for most agencies. Spending choices are subject to external evaluation so the government can verify public funds are well spent. In contrast, secretive intelligence agencies and certain parts of the military have access to unvouchered funds that allows them to spend money on research without explaining what it is for (Johnson, 2022, 168).[8] According to a senior GAO official, "we have no access to certain CIA 'unvouchered' accounts and cannot compel our access to foreign intelligence and counterintelligence information... [W]e have not actively audited the CIA since the early 1960s"(Hinton, 2001). Haines (1998, 85) notes that "scrutiny of the [intelligence] budget ranged between 'cursory and nonexistent.'"

One consequence of internal secrecy is that managers are *partly* forgiven for being ignorant when subordinates do things they do not expect them to do.[9] When a scandal erupts in an open government organization, a manager cannot easily say they did not know what their staff was doing because the public expects them to monitor employees. But national security employees are expected to maintain secrecy to guard against leaks and counter-intelligence threats. This helps excuse managers who do not intrusively monitor their staff to learn about questionable choices. During the Iran-Contra Affair, for instance, Reagan avoided some of the worst costs by claiming that subordinates engineered the scheme without his knowledge (Byrne, 2014).

We are interested in how internal secrecy impacts innovation in national security agencies. Our review suggests that secrecy is most salient during the early phases of innovation: periods where researchers develop prototypes or run laboratory tests and simulations without a manager or compliance officer knowing about it. Secondary accounts of DARPA program managers "start[ing],

[7] See Subpart 6.3, Federal Acquisition Regulation; https://www.acquisition.gov/far/subpart-6.3.
[8] See also Jacobsen (2015, 253).
[9] We parameterize the incomplete extent they are forgiven as x.

continu[ing], or stop[ping] research projects with little outside intervention" is a prime example (Jacobsen, 2015, 6). As projects progress, even secretive agencies may exploit open research practices to refine ideas by sharing information broadly across the national security community. But absent small teams pursuing initial testing in relative secrecy, many innovations may never make it that far.

To be clear, there are other parts of government with limited internal secrecy. For example, in parliamentary democracies, cabinet documents are sealed for decades so that elected leaders can brainstorm policy innovations (Cain, 2014). But this secrecy is confined to top-level policy discussion, and does not cover the design and testing of products. Pilot studies and focus group research to support policies formulated during cabinet are not privileged.

In practice, actors can exploit secrecy at different levels of a secret organization. To keep things simple, we detail a two-level institution that involves one decision-maker and one researcher. However, in many historical examples we see variation between who knows the devilish details and who does not. At one extreme, a handful of scientists know the controversial research activities but even their immediate superiors are unaware. At the other extreme, the executive is fully aware of the devilish details but legislators are not. In the middle, directors of intelligence agencies know exactly what their subordinates are doing but do not inform the executive.[10] If we add layers of management to the institution, our basic predictions still bear out so long as there is secrecy at some level of the organizational hierarchy. There must be at least one partition between insiders who can pursue research without explaining their practices outside of the group and who share the costs of authorization if things go wrong, and outsiders who can escape some costs by remaining ignorant about what her subordinates are up to but cannot stop programs for a long time.

3 Model

Our analysis plan is as follows. First, we set up a basic institution. Second, we formally define secret innovation and contrast the process of innovation in secret and open organizations to explain the core mechanism driving secret innovation. Third, we use comparative statics to explore the innovations uniquely pursued in secret organizations. Fourth, we introduce two distinct

[10] In other examples there are inter-agency teams. But the teams are small and secret. Our theory covers any project team that can maintain secrecy, whether all members work for the same agency or not.

information, agency, and monitoring problems to flesh out the mechanism and connect the model to the principal-agent literature. Finally, we consider the rationale for allowing secrecy given that it can lead to perverse outcomes.

3.1 Setup

We study an institution that employs two actors: a researcher (R, she) and a manager (D, for decider, he). Figure 1 visualizes the game-tree and payoffs. The dashed box is the sub-game in which R exploits secrecy. In it, she can conduct pilot research without her manager knowing about it. Below we contrast secret and open institutions. Open institutions remove the secret sub-game but are otherwise identical.

We model the true effect of unleashing a new innovation on the world as $\pi \in \mathcal{R}$. When π is positive (negative), it means the innovation ultimately moves the institution closer (further) from achieving its goals. Of course, actors cannot anticipate all the consequences of unleashing new devices *ex ante*. Thus, D's choice to innovate is based on an expectation of the consequences. Define $p(\pi) \rightarrow \mathcal{R}$ as a density function that determines the effect of introducing the innovation. We assume both players know the density function $p()$, but not the true realization of π.

Along the way to innovation, actors can authorize pilot research, which has two effects. First, pilot research improves the value of innovation by $\theta \geq 0$. Second, pilot research helps discover the true effect if innovation happens. We model this as a normally distributed signal $m \sim \mathcal{N}(\pi, \sigma)$ tied to the true consequences of innovation (π).[11]

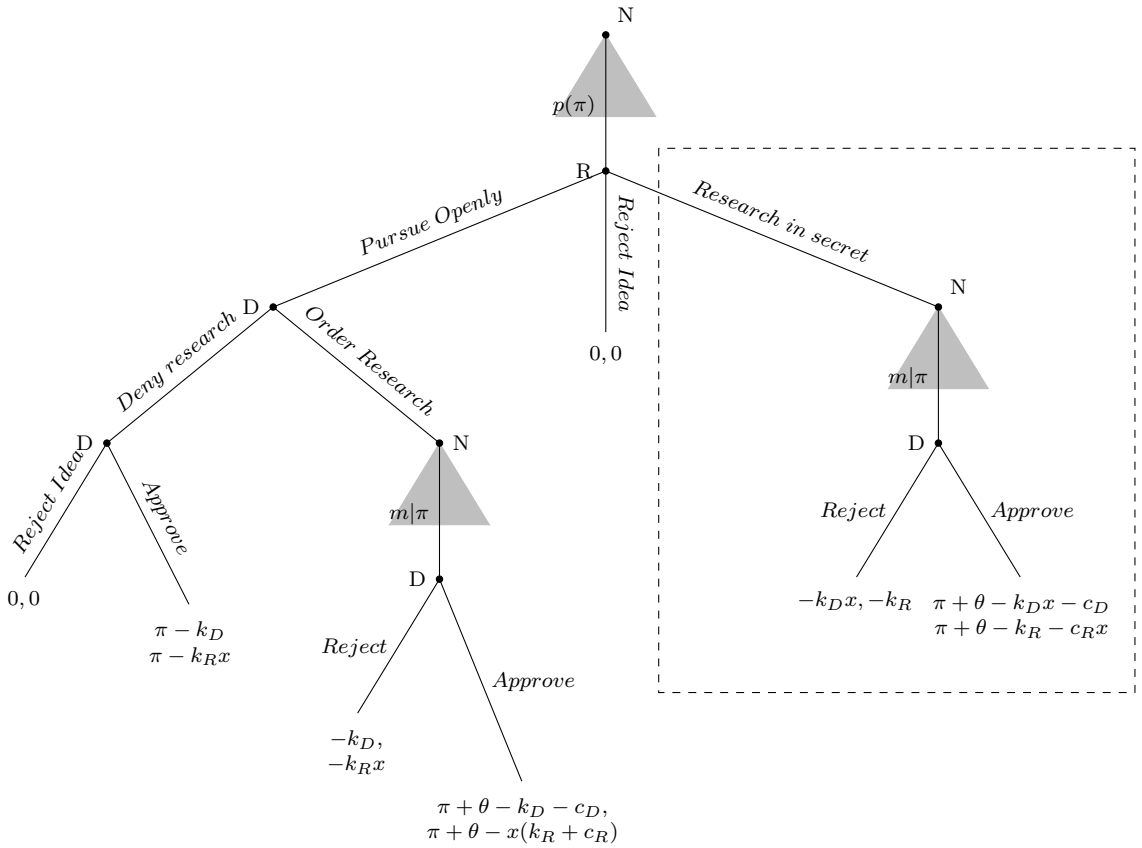
Actors pay political costs for participating in a controversial research process. We assume players pay one cost—$k_i, i \in \{R, D\}$— if the institution engages in pilot research.[12] They pay a second cost—c_i—if the project is deployed into the field. We assume that actors incur costs based on how responsible they are during the decision-making process. The total amount of cost to be apportioned in $1 + x$. We distribute 1 unit of cost to the actor that chooses to take costly action (conduct research, authorize innovation), and a smaller $x \in (0, 1)$ portion to the other actor who works at the same institution but did but did not directly take a costly action.[13]

[11] Note π is drawn from an arbitrary distribution. We model the signal from a normal distribution to avoid corners if $p(\pi)$ is supported on a limited range.

[12] A cost of $k = 0$ implies that pilot research isn't controversial.

[13] As is standard, c, k are an actor's cumulative expectation of harm incurred at the moment a choice is made based on the likelihood each possible punishment will be imposed, and agents' sensitivities to each punishment. x

Figure 1: Game Tree For Baseline Model



Dashed rectangle represents the secret option. The open organization omits this sub-game.
Shaded triangles represent random variables. Nature does not reveal π to either player. Nature reveals m to both players.

Parameter	Summarized Interpretation
$\pi \in \mathcal{R}$	Affect from unleashing an innovation on the world
$p(\pi)$	Initial expectation about the innovation's affect.
$c_i \geq 0$	Expected political costs actor i incurs when innovation is approved
$k_i \geq 0$	Expected political costs actor i incurs when pilot research is approved
$x \in [0, 1)$	What the actor pays who participates in an innovation phase that she did not authorize
$m()\|\pi$	What is learned about π from pilot research
$\theta \geq 0$	How much pilot research improves an idea

3.2 Analysis: Secret innovation and the cost-passing mechanism

Our solution concept in the main model, and unless otherwise states, is sub-game perfect equilibria (SPE). We define secret innovation as follows.

Definition Suppose parameter values in the open institution where innovation does not occur with probability on the path in any SPE. Then **secret research facilitates innovation** if innovation occurs with positive probability in any equilibrium in the secret institution with the same parameter

allows authorization and knowledge of wrong-doing to moderate this expectation.

values.

This definition highlights the counterfactual nature of our claim. Open institutions can innovate. But there are some ideas that only secret institutions will pursue.

Our first task is to identify the ideas open institutions will not pursue. There are two potential pathways. First, D can innovate absent research. Define, e_0 as the actors' prior expected value for π. Second, D can research and then decide to innovate if the research shows sufficient promise. We define two expectations at the moment D must authorize research (or not). Define $\lambda = pr(\mathbb{E}[\pi|m] > c_D - \theta)$. Informally, this is D's pre-research belief that if research is conducted, he will observe a signal m that will lead to a posterior belief that the project is sufficiently likely to hold benefits that outweigh the costs ($\mathbb{E}[\pi|m] > c_D - \theta$). That is, it is D's pre-research belief, that D will innovate after observing research. Define, $e_1 = \mathbb{E}[\mathbb{E}[\pi|m]|\mathbb{E}[\pi|m] > c_D - \theta]$. Informally, this is D's pre-research expected value of π, given that D will observe an m sufficiently large, that D is willing to approve research. Appendix A.1 reports more technical information on these expectations.

Lemma 3.1 *Neither research nor innovation can happen in the open institution if*

$$e_0 < k_D \tag{1}$$

$$\lambda < \frac{k_D}{e_1 + \theta - c_D} \tag{2}$$

In every SPE player utilities are $U^D = U^R = 0$.

See Appendix A.2. When condition 2 is satisfied, D conducts research to determine if the project is viable. Two factors drive D to reject a request for pilot research. First, research involves political costs (k).[14] Second, at the point where D is asked to authorize controversial research, his expectation about that research is inextricably connected to his prior belief. When preexisting scientific research suggests the project is not promising, D expects future research to, on average, confirm that expectation.

We now turn to the secret institution. Since we are interested in the cases where secrecy facilitates innovation, we focus on the conditions where innovation cannot happen in the open

[14] Trivially, if research is costless or beneficial you always see open innovation.

institution.

Proposition 3.2 ***Secrecy facilitates innovation*** *if conditions 1, 2 and:*

$$\frac{k_R}{e_1 + \theta - c_R x} < \lambda \tag{3}$$

are satisfied. If they are, then in every SPE, R exploits secrecy to conduct pilot research, D authorizes the project if and only if that research provides evidence the program will work. Off the path, if R attempts to pursue open research, D denies R's research and innovation does not happen.

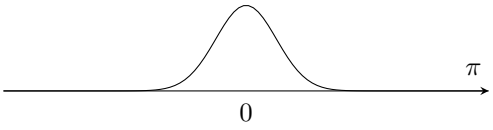
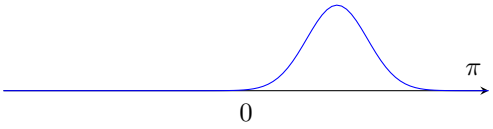
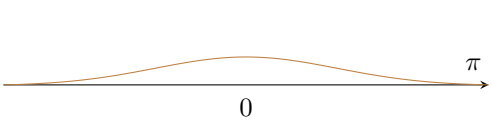
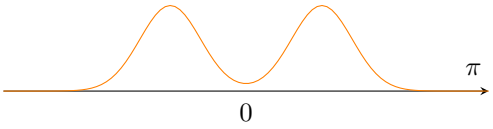
See Appendix A.3. The result describes a condition where the researcher is willing to exploit secrecy to conduct research (condition 3 is satisfied), but her manager was unwilling to approve open research (condition 2 is satisfied). If research provides evidence the project is viable (m suggests π is higher than originally thought), the manager will approve the project, leading to innovation.

Notice that we can achieve secret research even if the manager's and the researcher's cost parameters are identical: $k_R = k_D, c_R = c_D$. This is surprising given what we know about principal-agent problems. In standard accounts, researchers only exploit secrecy when their preferences diverge from the manager. Why is a researcher with the same incentives as the manager willing to conduct research when her manager is not? The answer comes down to cost passing. Secrecy gives the researcher discretion to conduct pilot research to try to convince the manager, who is unwilling to pay the research costs, to approve if it shows promise. If R's secret pilot research shows promise (m is large) she can take the results to her manager for approval. Thus, the researcher is willing to assume the up-front cost and risk of research because she can convince her manager to bear the brunt of deployment costs.

3.2.1 Predictions about ideas: Secrecy drives innovation when initial ideas are high-risk, high-reward

What are the kinds of initial ideas researchers need secrecy to pursue? Using a comparative static analysis, we expose two ideal-type pathways to secret innovation that are made possible because the manager and researcher weigh certain trade-offs differently. We provide technical support for these pathways in Appendix A.4. We visualize the results in Table 1. These pathways

Table 1: The innovation pathways for different initial ideas

Initial predictions of effects: $p(\pi)$	Substantive description	k_D low	k_D high
Baseline: $e_0 = 0.1, \sigma = 1$			
	Project team largely agrees innovation won't impact state-goals.	Scrap idea	Scrap idea
More Optimistic, same confidence: $e_0 = 3, \sigma = 1$			
	Project team largely agrees innovation will positively impact state-goals.	Open research	Secret research[2]
Very uncertain about consequences: $e_0 = 0.1, \sigma = 3$			
	Project team is widely uncertain about project effects.	Secret research[1]	Scrap idea
Foresee positive & negative consequences: $e_0 = 0.1, \sigma = 5$			
	Some predict huge success, others predict negative consequences.	Secret research[1]	Scrap idea

Rows represent different initial expectations about in innovation's effects ($p(\pi)$). Superscripts 1, 2 identify innovation pathways. Pathway 1 treats row 1 as the baseline, and raises variance (σ). Pathway 2, considers shift from low-to-high k_D across columns.

can interact. However, the basic trade-offs we identify are always present. Thus, it is valuable to consider them as distinct.

The first pathway appreciates the actor's initial expectations about whether an idea will provide a benefit ($p(\pi)$). In real life, a researcher uses publicly available research on related problems to make predictions about what will happen if her idea is developed. Column 1 of Table 1 plots the initial expected consequences of four different concepts institutions could pursue. Row 1 is the baseline. The remaining three panels represent different ways initial beliefs can vary.

First, they vary in their on-average, expected effects (e_0). As e_0 increases (row 2), it means the institution's initial expectation is that the idea is increasingly likely to yield a net benefit if it is developed and deployed into the field. The second way initial ideas vary is in the standard error of $p()$. We notate it σ. Substantively, a high standard error could represent two things. At the individual-level (row 3), it represents an idea that is so novel there is little else to compare it to. In

these cases, researchers do not know what to expect but accept that unleashing the idea on the world could have many unanticipated consequences. At the group-level (row 4) σ represents disagreement about the potential consequences of innovation. The debate surrounding autonomous weapons is instructive. Proponents emphasize greater speed and stealth with fewer casualties. Critics point out that they might create greater instability and more crises (Laird, 2020). Before these systems are deployed, it is hard to know if they will provide benefit or cause harm.

The following expectation summarizes one pathway to research under the assumption that the k_D is low (column 3):

> **Pathway 1: Deep uncertainty.** If the political cost associated with research is low, then secret research facilitates innovation if:
>
> - R is unsure if the innovation will yield benefits or costs once deployed ($e_0 \approx 0$). If instead she was confident that it will yield benefits $e_0 >> 0$ she would pursue open research.
> - The improvement value (θ) is not too large, and there is little preexisting scientific research. Therefore, the researcher is not confident in her initial expectation (σ is high). If instead she was more confident that she understood the idea's effects (σ was lower), she would scrap the idea.

Why does secrecy facilitate research when researchers are deeply uncertain about the project's effects? The logic relies on two steps. In Lemma 3.1 we showed that the manager only pursues research if her expected benefits for success are sufficiently high. Deep uncertainty implies an idea could generate large positive, or negative, effects. When D weighs these different outcomes his expectation of benefits is near 0. This is what we observe in rows, (a), (c) and (d) of Table 1. Of course, D could use research to learn more about whether the idea is viable. However, research is costly and D's expectation that pilot research will show promise is tied to his initial expectation of the innovation's effects (i.e. approximately 0).

In proposition 3.2 we showed that the researcher is also sensitive to expected benefits but is willing to pursue research under more conditions because she can distribute the costs. As a result, when the costs and expected benefits are both low, the researcher is willing to pursue secret research *so long as she believes her research will convince the manager to approve her idea.* The researcher believes that a manager is likely to be convinced by pilot research when σ is high.

One reason for this is that when there is little preexisting research, the researcher's pilot research

carries a larger weight in the manager's overall expectation of success. Another is that when projects are likely to have either extreme positive or negative consequences, pilot testing indicates which direction the program will go. If results are positive, D is confident the project will have major benefits and can accrue those by authorizing the project.

The second pathway to secret innovation relies on a trade-off between the political costs of research (k_D) and the expected consequences of deploying a new innovation (e_0). Substantively, k_D captures how sensitive the manager (and the institution at large) is to the expected moral and political costs associated with research when they authorize it.

> **Pathway 2: High stakes.** Secret research facilitates innovation when:
>
> - the expected benefit of innovation is high (e_0 is high); and
> - the manager's sensitivity to research costs are also high (k_D is high) but either the researcher's sensitivity is lower ($k_R << k_D$) or cost sharing is moderately calibrated to support proposition 3.2.
>
> If the manager's political costs of research and production were lower, we would observe open research.

The logic for the basic trade-off is simple. There are initial ideas that show enormous promise. However, the research required to pursue these ideas involves political costs. Secrecy facilities innovation when the manager is unwilling to bear the large costs of research and the costs of approval on his own. But once the research is complete, the manager will happily approve the project. In cases like this, the researcher may bear the 1 share of research costs knowing the manager will bear approval costs once research is complete.

3.2.2 Predictions about patterns of innovation: Secret institutions generate important innovations that open institutions do not

In terms of aggregate patterns of innovation, what are the features of research projects and innovations we expect from secret versus open institutions? We find that secrecy allows organizations to research ideas that seem bizarre, morally repugnant, or likely to fail when first conceived.[15] This leads to a straightforward expectation.

[15] For a long list of failed innovations, see (Houghton, 2019).

Expectation 1 *A larger proportion of ideas are rejected after secret research than after open research.*

We might intuit from this that secret innovation damages a nation's security in the aggregate. However, secret organizations are only willing to pursue these ideas because the potential upside is high. The initial idea must have a large enough chance of making a positive impact for a researcher to pursue it. If research confirms the idea is harmful, the institution scraps it early on. In the rare cases when research suggests an idea will provide benefits, these ideas are converted into innovations that change the world. This leads to a second prediction:

Expectation 2 ***Secret research leads to radical innovation.*** *Consider two comparable cases, a baseline case where R pursues open research because e_0, e_1 are sufficiently large, and a counterfactual case where R pursues secret research because the counterfactual values $e_{0\alpha}, e_{1\alpha}$ are smaller. Then so long as the true effect of innovation (π) is large, increasing the true effect of innovation further increases the chance innovation in the counterfactual case more than it does in the baseline case.*

There are two parts to this reasoning. First, in cases where managers approve open research, they are basically sold on the concept. Thus, even if the pilot shows only moderate success they will approve innovation. By contrast, in cases where researchers opt for secrecy, the manager starts out skeptical. Thus, the result of the pilot tests must be very strong to convince the manager to approve innovation. Second, the pilot test is correlated with the true effect. Thus, increasing the true effects has a greater impact on whether the message will induce secret research. This has an interesting empirical analog. For every handful of bizarre and shameful failed projects—bionic cat robots, nuclear-induced tsunamis, and so forth (see Houghton, 2019)—secret institutions provide a radical success—the reconnaissance satellite. With foresight, these innovations all sounded risky. With hindsight, some are radical innovations that shaped the industrial and digital revolution and medical sciences.

With foresight, these innovations all sounded risky. With hindsight, some are radical innovations that shaped the industrial and digital revolution and medical sciences.

3.3 Connecting the mechanism to the principal-agent literature

The basic model identified how secret innovation allowed actors to distribute costs at different stages of the innovation process so they could pursue a wider range of novel ideas. However, the model did not fully explore the perverse incentives that arise given uncertainty and principal-agent situations. We now introduce these into the model. We show that our basic logic survives and derive additional implications about how researchers and managers collaborate to exploit secrecy in national security institutions.

3.3.1 Monitoring

We assumed that if researchers exploit secrecy, the manager is forced to take on a $k_D x$ cost when the program comes to light. In practice, managers can monitor subordinates' activities by asking for project details. Given that the researcher's actions can force the manager to incur costs, why doesn't the manager monitor their activities?

This is at the heart of the principal-agent literature. Managers want to stop subordinates from taking actions they would not approve of. In this literature, there is agency loss because monitoring is difficult and expensive. However, if monitoring was costless and perfect, D would always monitor and R would always behave (Eisenhardt, 1989). This concern is relevant for our theory because R only uses secrecy because D will not approve.

We adjust the baseline model to capture monitoring as it is commonly studied in the principal-agent framework. First, we introduce uncertainty over research costs. We start with a simplifying assumption $k = k_R = k_D$. We then add a step at the beginning of the game where Nature selects the cost associated with research $k \sim f()$ where $f()$ is supported on the non-negative real numbers. Second, we assume that if the manager does not observe open research he has the opportunity to monitor the researcher's activities. If the manager chooses to monitor and discovers the researcher started a secret research program, he has two options: he can allow the research to continue or shut it down. If the manager allows the program to continue, the game reverts to open research (and associated payoffs). If the manager shuts the research down, the manager avoids research costs entirely, the researcher incurs k_R, and research has no effect (we do not realize θ, m).

We explicitly assume that D pays no cost to monitor, and if the manager does monitor, he

perfectly observes R's behavior. Indeed, this is the exact condition the principal-agent literature suggests should drive complete monitoring. Define $\bar{k} = \lambda[e_1 + \theta - c_R x]$, and $\underline{k} = \lambda[e_1 + \theta - c_D]$. Assume $0 < \underline{k} < \bar{k}$. Further define

$$\mathbb{E}[k|sr, nor] = \frac{\int_{\underline{k}}^{\bar{k}} k f(k) dk}{\int_{\underline{k}}^{\infty} f(k) dk}.$$

This represents D's expected cost k that D will incur if D fails to monitor, at the moment D must decide whether to monitor or not, and given D's expectation that he would not have observed research.

Proposition 3.3 ***The don't-ask-don't-tell equilibrium***. *Suppose conditions 1, 2 and 3 can be satisfied for some $k = k_R = k_D$, then in the model where D can perfectly monitor R, if*

$$\mathbb{E}[k|sr, nor] < \frac{\lambda[e_1 + \theta - c_D]}{x} \tag{4}$$

The following pure strategies are a Perfect Bayesian Equilibrium (PBE).

- *D does not monitor if research is unobserved. D approves open innovation iff $k < \underline{k}$ but not otherwise. Regardless of how research occurs, D approves post-research innovation if $\mathbb{E}[\pi|m] > c_D - \theta$. Off path, if D decides to monitor, D shuts-down research with a cost profile $k \geq \underline{k}$, then does not approve innovation. Also off-path, D rejects innovation absent research.*
- *R scraps the project if $k > \bar{k}$, R conducts open research if $k < \underline{k}$ and conducts secret research otherwise.*

Secrecy facilitates innovation if $k \in [\underline{k}, \bar{k}]$.

See Appendix A.6. This result is surprising. After all, the only reason the researcher does not ask for permission is that she knows the manager will not approve. Thus, when the manager observes the researcher hiding her activities, he should suspect something bad is happening and engage in monitoring. From the researcher's perspective, this is indeed what is going on: she is exploiting secrecy because she knows her manager will not approve her controversial research program. And yet, the manager elects not to monitor. Why? The logic follows a don't-ask-don't-tell dynamic made possible by cost passing. The manager knows if he monitors he will learn the

devilish details of what is happening and be forced to shut down the project, rendering a payoff of 0. However, if the manager does not monitor, he can reduce his costs through plausible deniability.

In this equilibrium, there are research protocols that are so controversial the manager does worse by allowing research to continue even though he only incurs an x share of the cost. Despite this extreme preference asymmetry, the equilibrium holds because the manager expects the researcher's protocol is too controversial to approve but not so controversial that the manager does not want the researcher to pursue it in secret. This leads to the following empirical implications:

Expectation 3 ***Don't-ask-don't-tell:*** *When managers are alerted that a researcher is secretly researching and does not want to share the details, they elect not to monitor because they suspect the program is controversial. Managers allows secret research to progress so they can retain plausible deniability.*

Expectation 4 ***Telling implies shut-down:*** *If managers observe controversial details of a research program that a researcher secretly pursued, they shut down the parts of the program they observe.*

3.3.2 Trust when the researcher can fabricate her report

The analysis above emphasized that secrecy has positive effects because it provides researchers with autonomy; managers with cover from political costs; and both actors the capacity to distribute costs between them. In practice, secrecy also creates opportunities for R to fabricate reports or cherry-pick results. In theory, it could cause the entire secret research program to unravel. Secret research only works if the manager can trust the researcher's description of pilot results.

We adjust the baseline model to understand if the manager can assign a researcher to a project who will pursue controversial pilot research if it is necessary, and credibly reveal the results of that pilot. First, we assume that if research is conducted in secret, only the researcher observes m. Second, we assume the researcher can write any (costless) report she likes: $m_r \rightarrow \mathcal{R}$.[16] When research happens in secret, the manager only observes the report m_r. We say the research report is honest if $m_R = m$ and dishonest otherwise. Third, we allow D to set the researcher's cost profile c_R, k_R, which represents a manager's ability to assign projects to staff. In short, we want to know

[16] Trivially, adding dishonesty costs makes honesty easier.

if managers can find a researcher who (1) is willing to conduct secret research; (2) is willing to write an honest report no matter the outcome of her pilot; and (3) that the manager will believe. Finally, we want to know if the Manager would like to employ a researcher who embarks down secret research.

Lemma 3.4 *If conditions 1-3 and*

$$\lambda > \frac{k_D x}{e_1 + \theta - c_D} \tag{5}$$

are satisfied then the manager employs a research that is honest, trustworthy, and willing to conduct secret research.

See Appendix A.7 for a technical statement of lemma 3.4 and proof. Lemma 3.4 explains that it is possible to find a researcher who can facilitate secret innovation. But what does this researcher look like? We put the answer in terms of expectations:

Expectation 5 ***Secret research only works if the institution employs unscrupulous patriots****. The researcher that takes on a secret research program and will report her results credibly and honestly must be:*

- *insensitive to the political and moral issues associated with research ($k_R \to 0$), but*
- *highly sensitive to the foreign policy costs associated with deploying a project ($c_R = c_D/x$).*

The first bullet-point summarizes the condition where the researcher is willing to pay the cost to conduct controversial pilot research even if the manager is not. The second summarizes what it takes for the researcher to honestly report pilot research. To be clear, the condition on c_R for complete revelation is a strict equality that aligns R and D's preferences at the point where D must decide between approving innovation or not. However, we can still support the credible revelation of information, with honesty in some cases and dishonesty in others, with some cost asymmetry. For example, there are cases where the researcher is less sensitive to the costs of deployment ($c_R < c_K/x$) where D is still persuaded by R's research report and innovates if R's report is positive. In this case, it is possible R wants to innovate following pilot research, but D would not innovate if he knew the truth. In these cases, R fabricates the report. Had R sent an

honest report, D would have rejected it. D is aware of this risk but trusts R anyway because the results of pilot research that generate incentives for dishonesty are unlikely relative to the results of pilot research where both actors would proceed.

In short, D will trust R even if their preferences are not perfectly aligned because R is sufficiently sensitive to the foreign policy costs associated with innovation so that R does not want to approve projects that are likely to fail in most cases. This result has a secondary implication about how a researcher who has selected into secret research will behave following the outcome of pilot research:

Expectation 6 *Suppose a researcher is willing to take on a secret research project. Then, if pilot research suggests that a project will fail, the researcher will terminate the research and argue against developing the project because she believes the chance of failure is high.*

3.4 External ambiguity and calibrating cost passing

Because secrecy makes oversight hard, managers could sustain plausible deniability over the devilish details if the researcher briefed the manager informally. This would facilitate oversight while offsetting the manager's expectation of incurring the increased costs from authorization should the controversial aspects ever be exposed. Even if managers learn informally and passively approve, though, they are still more exposed to costs in expectation than if they learned nothing. For example, if a controversial experiment comes to light, an investigator may piece together the manager's knowledge from unusually long meetings with the research team, coded messages, or depositions of subordinates. Thus, at the time the manager is informally briefed, the decision to approve research must still factor in the cost of professional disgrace and criminal liability from involvement (k_D)—and the expectation of incurring these costs given the raised expectation of incurring these costs from the informal briefing (call it, $x + z < 1$). Here, the expectation is lower than if the manager had written a memo authorizing the experiments but higher than if they were truly ignorant (x).

In Appendix A.8, we extend the model to account for these issues. We set up the model as a tough test for internal secrecy because the researcher faces strong incentives to brief informally to pass on at least some costs, and we assert that if the researcher does so it does not meet our definition of internal secrecy.[17] And yet, we still find that researchers exploit internal secrecy (i.e.

[17] There is still an indirect effect of internal secrecy in that the manager's ability to offset costs comes from the fact that an unmodeled higher-order principal cannot observe informal manager-researcher interaction. We also discuss

do not brief the manager at all) rather than provide informal briefings when the underlying costs parameters (k_i, c_i) are high. What is more, we show that the option to brief informally raises the chance that research occurs beyond the baseline model. This illustrates how modeling other loose reporting requirements that internal secrecy facilitates expands the conditions under which innovation occurs.

3.5 Institutional design

In theory, even the president, the most senior member of the executive, answers to Congress (and the public). If abuse is possible, why does Congress tolerate the institutional arrangement described above? Why doesn't Congress design institutions that hold the executive accountable even when they do not learn the details? It is hard to address this question empirically because internal secrecy is an enduring feature of national security institutions. In the US context, the National Security Act of 1947 handed the executive and national security agencies enormous power to sustain internal secrecy (Byrne, 2014, xv). This legislative framework survived reform debates that followed intelligence failures and executive abuse.

One possibility is that reform is hard and institutions are sticky. But in Appendix A.9, we adapt the monitoring model to provide a strategic explanation for Congressional inaction. We introduce a higher-order principal (Congress) who first sets $x \in [0, 1]$ (the level of internal secrecy), then the game unfolds as in section 3.3.1 given the x Congress has set. Our set-up closely reflects two features of Congress' abilities and incentives expressed during historical debates over reform. First, Congress's main power to influence national security employees is by passing *ex-ante*, rather than scandal-specific, laws about appropriate conduct for all future cases of secret research. This includes when managers are supposed to monitor their subordinates, when subordinates must report their activities, and so on. Then, national security employees are confronted with specific scenarios (e.g. the decision to pursue a particular idea) knowing the laws that govern their actions. Second, Congress is aware that internal secrecy is necessary to sustain external secrecy. Others have shown that greater oversight, or even greater sharing within the national security community, runs the risk that foreign threats will learn about sensitive operations (Joseph et al., 2022). Thus, Congress knows the higher it sets x, the more likely it is US rivals will discover secrets and exploit them.

another model where informal briefings can sustain internal secrecy, that yields stronger results in favor of our theory.

We focus on conditions where, as shown in Section 3.3.1, if x is sufficiently low Congress induces the researcher and manager to engage in the behaviors described in the don't-ask-don't tell equilibrium. But if Congress sets x higher, they induce the researcher to never engage in secret research and we only observe non-controversial research the manager directly approves. Our model provides two strategic explanations for why Congress sets x low. The first closely reflects the don't-ask-don't tell mechanism. Congress also desires welfare-enhancing innovations, and knows innovation is less likely if x is high. When the costs/risk of abuse are low relative to the foreign policy stakes, Congress prefers to tolerate a risk of abuse for the same reason the manager prefers not to monitor. Second, when the trade-off between internal and external secrecy is severe, Congress prefers to tolerate the risk of abuse to prevent foreign threats from discovering secrets. This second mechanism potentially explains the unique amount of internal secrecy in national security agencies. For example, there is little cost of leaking innovations in education policy because they will not be exploited by rivals. Thus, Congress holds no incentive to write laws maximizing internal secrecy. Concerns over national security leaks can cause Congress to tolerate the risk of abuse from internal secrecy in national security institutions. Radical innovation is a convenient byproduct.

4 Testing the Argument

We trace the logic of secret innovation in two cases: the search for mind control (MKULTRA) and the first reconnaissance satellite (CORONA). Table 2 summarizes the case parameters, which we substantiate in the sections below. As a reminder, our theory identifies two pathways to secret research. MKULTRA fits our high risk-high reward pathway. Its moral repugnance generated enormous political costs during the research phase. But the promise of mind control was seen as a major benefit. CORONA fits our lower cost but high variance pathway. The political costs from CORONA are smaller because they stemmed mainly from perceptions of wasteful spending. But so little was known about the atmosphere and satellite telemetry that researchers found it hard to predict its chance of success.

Table 2: Summary of case coding

MKULTRA	
Research Team (R)	Sidney Gottlieb; Richard Lashbrook
Managers (D)	CIA Director Allen Dulles; Assistant Deputy Director of Plans Richard Helms
Anticipated Cost When Research Approved (k_i is high)	Research program involved experiments on US citizens without their knowledge/consent. If the project was exposed those who knew these controversial aspects could face prosecution, professional disgrace, job loss, etc. Whether exposed or not, actors could be psychological sensitive to inflicting harm on innocent civilians.
Apportioning Costs $(x, 1)$	Prosecution and professional disgrace were most likely for those who ordered controversial aspects. Psychological harm only applies to those who know.
Distribution of Expected Benefits ($p(\pi)$ reflects Table 1, row 2.)	Actors were cautiously optimistic that mind control could be achieved, They thought the national security benefits were substantial if it was, and saw little chance an effective mind control policy would hold disutility.

CORONA	
Research Team (R)	USAF Major General Bernard Schriever; USAF Col Frederic Oder
Managers (D)	Secretary of the Air Force Donald Quarles
Expected Costs (k_i is low)	Executive may perceive those who authorize as undermining Eisenhower's space for peace policy; and may be perceived as wasteful stewards of public funds at a time of tight budgets.
Apportioning Costs $(x, 1)$	Quarles knew he was closely monitored by Congress and would bear responsibility if he approved.
Distribution of Expected Benefits ($p(\pi)$ reflects Table 1, row 3.)	Many believed satellites were unviable. Some concerned that if they worked they could cause escalation/crises. Others were hopeful.
Institutional Quirk	While funding CORONOA was internally open at USAF, CIA offered an internally secretive funding vehicle. See text for details.

4.1 Mind control

In the late-1940s and early 1950s, US policymakers became convinced that the Soviet Union and the People's Republic of China had mastered mind control (Thomas, 1989, 94).[18] According to Richard Helms, a longtime CIA official who would go on to become Director: "There was deep concern over the issue of brainwashing... We felt that it was our responsibility not to lag behind the Russians or the Chinese in this field." (Kinzer, 2019, 54).

Policymakers were hopeful they could unlock the mysteries for themselves (U.S. Senate, 1976, 385). They believed mind control was "of the utmost importance... [and] could mean the difference between the survival and extinction of the United States" (Kinzer, 2019, 49). A declassified memo from the early 1950s list core aims: "A. Can accurate information be obtained from willing or unwilling individuals. B. Can Agency personnel... be conditioned to prevent any outside power from obtaining information from them by any known means? C. Can we obtain control of the future activities (physical and mental) of any given individual...?" (Redacted, 1952, 1).[19]

In 1950, the CIA conducted some ad-hoc experiments codenamed BLUEBIRD, then ARTICHOKE (McCroy, 2006, 26-27).[20] Even these initial projects were handled outside normal oversight channels. A memo to the CIA Director stated: "In view of the extreme sensitivity of this project and its covert nature, it is deemed advisable to submit this project directly to you, rather than through the channel of the Projects Review Committee. Knowledge of this project should be restricted to the absolute minimum number of persons" (CIA, 1950, 1).

Within a few years CIA decided to "intensify and systematize" their efforts. In April 1953, CIA Director Allen Dulles authorized Sidney Gottlieb to establish MKULTRA (Kinzer, 2019, 72-73). Gottlieb was allowed to conduct experiments with virtually no oversight. Years of controversial experiments followed. Consistent with our assumptions, CIA managers granted this level of secrecy to researchers partly because of external threats. The Technical Services Division was awarded "exclusive control of the administration, records, and financial accountings of the program" owing to fear that "Public disclosure of some aspects of MKULTRA activity could ... stimulate offensive and defensive action in this field on the part of foreign intelligence services" (Earman, 1963, 2).[21]

[18] See CIA (1956).
[19] See also CREST (2011, 2).
[20] See also Streatfield (2007, 27).
[21] They also worried that public disclosure "could induce serious adverse reaction in US public opinion." See

While Dulles gave the research team broad authority to conduct experiments involving "the use of biological and chemical materials in altering human behavior" (Earman, 1963, 30), he and other managers[22] were not privy to the controversial details of how this research was performed.[23] Gottlieb secretly tested the effects of LSD on unwitting, non-volunteer subjects (U.S. Senate, 1976, 391-392). Under Operation Midnight Climax, sex workers lured unsuspecting American citizens to a safehouse in San Francisco where CIA staff secretly administered LSD and monitored them (Kinzer, 2019, 141-152).[24] MKULTRA also involved experiments on prisoners overseas (Kinzer, 2019, 106). When the Church Committee reviewed MKULTRA years later, it was these research practices that caused them to conclude that "the nature of the tests, their scale, and the fact that they were continued for years after the danger of surreptitious administration of LSD to unwitting individuals was known, demonstrate a fundamental disregard for the value of human life" (U.S. Senate, 1976, 386).

As we demonstrate below, this firewall between managers and researchers meant the latter, who oversaw the experiments, were at greatest risk for potential criminal prosecution and professional disgrace. CIA managers, who were ignorant of the most controversial aspects of MKULTRA, suffered costs to a lesser degree.

In summary, several features of this case fit our high-stakes pathway for secret innovation. It involves two primary actors: the MKULTRA research team (with Gottlieb at the center), and CIA management (the most senior was Dulles). At the outset, Dulles knew that if MKULTRA succeeded, it would generate large benefits (e_0 was high). However, he also knew the necessary research would be controversial (c_i was high).[25] Starting from this position, three facts about this case match the

Earman (1963, 2).

[22]Richard Helms, Assistant Deputy Director for Plans, sat between Dulles and Gottlieb in the institutional hierarchy. We code him as a manager. The declassified record suggests he knew more than Dulles, but how much more is unclear. For example, in May 1953 Helms called LSD "dynamite" and said he "should be advised at all times when it was intended to use it." But he appears not to have been aware of some of the most egregious experiments (U.S. Senate, 1976, 395-96). Moreover, evidence of Helms advocating for unwitting testing is clearest in 1963—as the program was being shut down (U.S. Senate, 1976, 394). As Kinzer (2019, 154) notes, "Only two officers—Gottlieb and Lashbrook—knew precisely what it was doing." Ultimately, this is not especially consequential for our analysis. Our theory works in tiered institutions. The case would thus still fit if Helms was informed of some but not all of what Gottlieb did.

[23]Those above Dulles knew even less. As the Church Committed noted, "there were no attempts to secure approval for the most controversial aspects of these programs from the executive branch or Congress." See (U.S. Senate, 1976, 394).

[24]See also NBC (1977).

[25]To be clear, they did not know how controversial. The Inspector General who audited MKULTRA similarly acknowledged these trade-offs. See Streatfield (2007, 87).

choices our model predicts. First, the CIA hand-selected Gottlieb to oversee MKULTRA. Second, Gottlieb assessed that highly controversial human subjects research was necessary for MKULTRA. He could have discussed these research plans with managers but chose to keep these details secret. Third, Dulles had several opportunities to learn what Gottlieb was up to but never asked.

4.1.1 Why was Gottlieb chosen?

Gottlieb was not an obvious pick to lead MKULTRA. Although he had experience in government laboratories as a chemist, he did not have an intelligence background. Why was an intelligence outsider selected to lead a high-stakes and intensely secret project? In extension 3.3.2, we argued that when researchers conduct scientific tests in secret, it is easy for them to give managers the mistaken impression that their novel idea is more effective than the research suggests. Anticipating this problem, the manager must carefully select an unscrupulous patriot: a researcher who is insensitive to whatever controversy it takes to complete a research program, but who shares the manager's desire to only field projects that will advance national interests.

According to secondary accounts, this is exactly how CIA managers saw Gottlieb and others on the Technical Services Staff. The Agency needed "a character steely enough to direct experiments that might challenge the conscience of other scientists, and a willingness to ignore legal niceties in the service of national security' " (Kinzer, 2019, 47). The problem in Dulles' view was that certain parts of the CIA "had shown no stomach for further work on humans." As Thomas (1989, 98) notes, however, "the Agency's Office of Technical Services (TSS) had no such qualms... They would have no reservations about testing ideas on unsuspecting subjects, especially in such a vitally important and urgent area as brainwashing."

Accounts of Gottlieb's personality in particular are telling. According to Kinzer (2019, 50), "Like many Americans of his generation, he had been shaped by the trauma of World War II [which] left him with a store of pent-up patriotic fervor. His focused energy fit well with the compulsive activism and ethical elasticity that shaped the officers of the early CIA." When he later testified before a Senate Subcommittee about MKULTRA, Gottlieb used language we would expect from unscrupulous patriots: "I would like this committee to know that I considered all this work ... to be extremely unpleasant, extremely difficult, extremely sensitive, but above all to be very urgent and important ... [T]here was a real possibility that potential enemies... possessed capabilities in

this field that we know nothing about, and the possession of those capabilities... combined with our own ignorance about it, seemed to us to pose a threat of the magnitude of national survival" (Kinzer, 2019, 238).

Of course, Gottlieb faced incentives to cast himself as patriotic during an inquiry into his conduct. However, his behavior in the final years of MKULTRA also fit this personality profile. We show the patriotic researcher only pursues her project because she believes the science is viable. If she learns her research will fail to advance national security interests, she will quit even if no one stops her. Consistent with this logic, one reason why key parts of MKULTRA ended after nearly a decade of experimentation was Gottlieb's realization that "On the scientific side, it has become very clear that these materials and techniques are too unpredictable in their effect on individual human beings, under specific circumstances, to be operationally useful" (Kinzer, 2019, 198). It would be curious for a researcher motivated by pride to publicly declare their work a failure.

4.1.2 Why did researchers opt for internal secrecy?

If our theory is correct, Gottlieb and his team exploited internal secrecy because they knew CIA managers would refuse to let them continue the most controversial experiments if they figured out what they were up to. Unfortunately, Gottlieb never explicitly articulated why he kept the most controversial details of experiments from his managers. But the context surrounding his actions is consistent with our logic in three ways.

First, the experiments he was engaged in, particularly the parts having to do with surreptitious testing of unwitting subjects, were extraordinarily controversial. According to the Inspector General's report in 1963, "Research in the manipulation of human behavior is considered by many authorities in medicine and related fields to be professionally unethical, therefore the reputations of professional participants in the MKULTRA program are on occasion in jeopardy." It also states that "Some MKULTRA activities raise[d] questions of legality implicit in the original charter" (Earman, 1963, 1-2).[26] A memo from the late-1950s entitled "Influencing Human Behavior" similarly notes that "some of the activities are considered to be professionally unethical and in some instances border on the illegal" (quoted in Streatfield (2007, 86).) Because of this, "CIA officers

[26]The original charter here refers to the memo Dulles wrote authorizing MKULTRA. This further evidences that Gottlieb undertook activities that were illegal unbeknownst to Dulles.

felt it necessary to keep details of the project restricted to an absolute minimum number of people" (U.S. Senate, 1976, 406).

Second, several CIA managers stated they would have stopped MKULTRA if they had known about its full extent. According to Thomas (1989, 100), Dulles was interested in trying "everything the Communists could have done" but knew that "The risks for him and the Agency were enormous. If it ever became known that the United States government had funded what would be unprecedented clinical trials—ones beyond all ethical acceptability—it would most certainly lead to the sudden end of his remarkable and brilliant career." This is likely why, as detailed in the next section, he was cut out of the loop of the precise details of MKULTRA. The Executive Director-Comptroller, who was "excluded from regular reviews of the project," was strongly opposed to MKULTRA—when he learned about it. According to one account, "it is possible that the project would have been terminated in 1957 if it had been called to his attention when he then served as Inspector General" (U.S. Senate, 1976, 409).

Although less directly relevant given timing, Stansfield Turner, who served as CIA Director in the late-1970s, echoed similar considerations: "It is totally abhorrent to me to think of using a human being as a guinea pig... I am not here to pass judgment on my predecessors, but I can assure you that this is totally beyond the pale of my contemplation of activities that the CIA or any other of our intelligence agencies should undertake" (Kinzer, 2019, 234).

A final piece of evidence supporting the notion that internal secrecy facilitated Gottlieb's experiments is the fact that once Congress got wind of MKULTRA and asked to review the program files, Gottlieb destroyed them "on the verbal orders of DCI Helms" rather than handing them over (U.S. Senate, 1976, 403-404).[27] The destruction of records impeded subsequent investigations into the details of what transpired (Maret, 2018, 29). Gottlieb and Helms purportedly felt that the experiments "might be 'midunderstood'," leading them to direct "that every scrap of paper relating to the brainwashing experiments be incincerated" (Streatfield, 2007, 332).

4.1.3 Managers built the system so they were in the dark

Our theory suggests that managers will embrace ignorance because they know that if they do not investigate, they will incur a small cost as an ignorant bystander. But they may accrue a large

[27] See also CREST (2011).

gain from a successful innovation. If they investigate, they are faced with the choice of incurring a large cost or shutting down the program altogether. Three case features support this logic.

First, Dulles minimized his exposure to MKULTRA's details from the outset (Maret, 2018, 47). When he initially authorized the project in 1953, the $300,000 he set aside was "not subject to financial controls" and researchers had "permission to launch research and conduct experiments at will" (Kinzer, 2019, 73). Dulles' 1953 memo states that "The nature of the research and the security considerations involved preclude handling the projects by means of the usual contractual arrangements" (Dulles, 1953, 1). According to one account, "Dulles ordered the Agency's bookkeepers to pay the costs blindly on the signatures of Sid Gottlieb and Willis Gibbons, a former US Rubber executive who headed TSS" (Marks, 1979, 57). Helms, who was one of the few senior officials to have reasonable insight into MKULTRA, "avoid[ed] oversight even by the CIA's director, because he 'felt it necessary to keep the details of the project restricted to an absolute minimum number of people"' (McCroy, 2006, 28). Richard Lashbrook, one of the senior scientists alongside Gottlieb, purportedly stated at one point that "what was actually signed-off on was not the same as the actual proposal, or actual detailed project" (quoted in Maret (2018)).

Second, CIA managers went to great lengths to avoid looking into MKULTRA. The most extreme example involved a civilian employee of the Army, Frank Olson, who was unwittingly given LSD and purportedly jumped out of a hotel window to his death in the weeks afterwards. The internal investigation that followed accused the TSS of "fail[ing] to observe normal and reasonable precautions." In response, Dulles wrote a letter to Gottlieb "criticizing him for 'poor judgment... in authorizing the use of this drug on such an unwitting basis and without proximate medical safeguards' " (U.S. Senate, 1976, 398). Ultimately, however, these were not formal reprimands, had no effect on advancement, and did not lead to a termination of the experiments (McCroy, 2006, 30). Surprisingly, but consistent with our theory, even after investigators uncovered wrongdoing in the narrow experiments related to Olson, they did not expand their audit to MKULTRA broadly. According to one account, a senior CIA official cautioned that a formal reprimand "would hinder 'the spirit of initiative and enthusiasm so necessary in our work' " (Marks, 1979, 84).

Third, when MKULTRA was eventually made public, the costs were distributed in accordance with our theory. As the most senior scientist who knew the complete details, Gottlieb was hauled before Congress to testify. Years later, he was implicated in a variety of lawsuits of families of

victims of MKULTRA. Most important for our purposes, "since Richard Helms was not alleged to have been directly involved in the drugging, he could not be prosecuted—but ... the case against Gottlieb could proceed" (Kinzer, 2019, 256-257).[28]

4.2 Overhead reconnaissance

We chose the origins of the first US reconnaissance satellite (CORONA) for three reasons. First, it verifies that our argument extends beyond morally repugnant programs, like MKULTRA, to the costs and risks faced by many technical innovations. Second, reconnaissance satellites are a tough technological test of our theory because they are hard to keep secret; and because the research needs for cutting-edge experts across many scientific areas made openness attractive. Finally, there are historical quirks that provide a quasi-counterfactual test. CORONA occurred in a unique period in which the CIA was not widely known to be in the business of technical intelligence. Because of this, we know what *would have* happened if an open organization—i.e. the Air Force, where it was originally pitched—was the only avenue for authorizing this bold innovation; it was rejected.

4.2.1 The open origins of CORONA

Monitoring the Soviet Union was a pressing issue for policymakers in the early Cold War (May, 1998, 21).[29] As Soviet capabilities to thwart existing reconnaissance tools advanced, concerns about the continued viability of the U-2 spy plane grew. US policymakers wanted a more reliable option (Greer, 1973, 3). Thus, some in the Air Force conceived of Weapons System 117L (the antecedent to CORONA) (Brugioni, 2010, 200). Responsibility for it was placed in the Western Development Division, which was managing ballistic missile development. According to a declassified history, "WDD had been established with handpicked military personnel and with special reporting channels for expediting program decisions" (Oder et al., 1988, 4). They initially solicited design bids from cleared government contractors. Lockheed subsequently won a contract, but funding challenges loomed (Dienesch, 2016, 129).

The institutional structure surrounding WS-117L was internally open. The Secretary of the Air Force, Donald Quarles, "responded to news of the [Lockheed] contract by ruling that neither

[28] See also Calabresi, Cabranes and Heaney (1998).
[29] US interest in military satellites emerged circa 1945 (p30 Wheelon, 1998).

mockups nor experimental vehicles should be built without his specific prior approval" (Oder et al., 1988, 5). In other words, the research team lacked the ability to pursue pilot testing without alerting their manager. Moreover, although WS-117L was technically a classified project, presumably to keep information from the Soviet Union, "Program details were reported to, and approved by, Congress" (Oder et al., 1988, 14).

From the perspective of Air Force managers, approving research into WS-117L presented low (but non-zero) political costs, but uncertain *expected* benefits. On the latter, there was deep uncertainty about whether satellites were viable despite their enormous potential. Per a declassified history, "The technology to be embodied in the WS117L satellite was largely unproven; no satellite had even been orbited, and little was known of problems that might arise in a weightless, airless environment." It also notes that Quarles "was not actively hostile to the satellite program as such, but had developed strong views about reliability and using low-risk technology" (Oder et al., 1988, 6-7). Additionally, there was concern about unanticipated escalation. On the costs side, Eisenhower was promoting the "space for peace" initiative which had "become a credo of US policy in 1955" (Oder et al., 1988, 5). Decision-makers worried that if they authorized WS-117L, they would be perceived as acting contrary to such commitments. Further, WS-117L was so novel that research into it could be perceived as wasteful. Quarles understood "the administration's commitment to eliminate 'noncritical' defense expenditures." Weighing these costs and benefits, and despite the desire of the WS-117L research team, he "found ample justification for his stubborn refusal to approve the start of a meaningful development program (Oder et al., 1988, 6-7)."

After it became clear that Air Force management would not adequately fund WS-117L, a plan was hatched to pursue it secretly. The concept, conceived by Colonel Oder, was known as "Second Story" (Dienesch, 2016, 131-134). It had two prongs. First, it would be announced that WS-117L was being cancelled and replaced with a scientific satellite overseen by the Air Force. This was a cover story. At the same time, the project would be covertly restarted and accelerated under the auspices of the CIA (Oder et al., 1988, 10). As noted, the CIA was just getting into the business of technical intelligence and thus was not an obvious choice to handle the project. This is likely why it did not originate there. Interestingly, however, a handful of the individuals involved with WS-117L were familiar with the Office of Science and Technology after working on the highly-classified U-2 project (Richelson, 2002, 23). Thus, the very fact that they proposed this option, which was outside

of the " 'normal' development cycle" (Oder et al., 1988, 9) is highly suggestive that internal secrecy was viewed, at least by the research team, as a way to advance a bold and risky innovation.[30]

Sputnik's success in October 1957 took policymakers by surprise. While earlier behavior was obviously not conditioned by an event which had not yet taken place, the Soviet Union's success in space altered their thinking, including on the importance and feasibility of this technology (Wheelon, 1998, 32). As such, the post-Sputnik period is effectively a separate case and beyond our current scope. Moreover, policymakers' emphasis on limiting many discussions to oral briefings "owing to the extreme sensitivity" on the project means that "there are few official records in the project files bearing dates between 5 December 1957 and 28 February 1958" (Oder et al., 1988, 15) Nevertheless, our theory illuminates several key elements of this period that are worth highlighting.

First, the strong desire for external secrecy—in this case, concealing CORONA from the Soviets—meant that the CIA's ability to "maintain effective secrecy" was of paramount importance (Brugioni, 2010, 200-201).[31] Second, the value of preserving external secrecy resulted in deep internal secrecy, as evidenced by Eisenhower's admonition that "only a handful of people should know anything at all about it" (Oder et al., 1988, 20). The fact that the CIA Director was "the only US Government employee authorized to spend money without substantiating vouchers" is also notable in that it almost certainly helped avoid scrutiny of higher-order principals like Congress from interfering (Oder et al., 1988, 21). Eisenhower's apparent decision to approve CORONA via "a handwritten note on the back of an envelope," combined with the heavy emphasis on oral briefings, is also consistent with our mechanism focused on plausible deniability (Oder et al., 1988, 28).[32]

5 Conclusion

We argued that secretive national security institutions are more innovative *because* they are secret. Secrecy is not equally valuable at every stage of innovation. Rather it allows an enterprising researcher to pursue initial ideas that are so bizarre, morally controversial, or unlikely to work *ex ante* that their manager would refuse to fund the initial concept. But if pilot research confirms the researcher's intuition, she can convert it into an innovation. These ideas reflect some of the

[30] Initially, Second Story was "entirely concocted within Schriever's own organization." See Oder et al. (1988, 12).

[31] See also Wheelon (1998, 33). Consistent with our internal secrecy logic, officials cite domestic concerns as one major justification for bringing testing to CIA.

[32] Some declassified CORONA documents still have redacted dollar amounts. See CIA (1958, 5)

most important innovations of the last century. The model explains that this theoretically drives different patterns of innovation in national security and other public-sector agencies.

While we emphasized the welfare enhancing effects of internal secrecy, our framework is general. Future researchers should explore the the promise and peril of secrecy for innovation. They could consider how other institutional features could maximize innovation while reducing the risk of abuse. Additionally, they could examine diversity in institutional design to harness the late-stage advantage of open organizations and the early-stage advantages of secrecy.

These insights also have significant policy implications, particularly as it relates to the return of great power competition generally and competition between the US and China specifically. On the one hand, innovation is viewed as a key pillar of this dynamic. Mike Rogers and Glenn Nye, two former US representatives on opposite ends of the political spectrum, argue in an op-ed that "The race to take leadership in advanced technologies such as artificial intelligence, quantum computing, and 5G networks will determine the future balance of geopolitical power" (Rogers and Nye, 2019). In addition to big and bold innovation, officials have also emphasized political and ideological factors as relevant to great power competition. The Biden administration's National Security Strategy makes frequent mention of transparency and openness as being integral to competing with opaque, closed states like China and Russia (Biden, 2022).

Our framework and findings suggest that there is a potential tension between these two impulses. In particular, internal secrecy—which sits uncomfortably alongside calls for greater openness domestically and internationally—has facilitated some of the most radical innovations of the last century. Ultimately, the best course of action may be to maintain a diversity of institutions.

References

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith and Peter Howitt. 2005. "Competition and Innovation: An Inverted-U Relationship*." *Quarterly Journal of Economics* 120:701–728.

Biddle, Stephen, Julia Macdonald and Ryan Baker. 2018. "Small footprint, small payoff: The military effectiveness of security force assistance." *Journal of Strategic Studies* 41:89–142.

Biden, Joseph R. 2022. "National Security Strategy." *The White House* .

Boushey, Graeme. 2016. Targeted for diffusion? How the use and acceptance of stereotypes shape the diffusion of criminal justice policy innovations in the American States. Vol. 110 Cambridge University Press pp. 198–214.

Brugioni, Dino A. 2010. *Eyes in the Sky: Eisenhower, the CIA, and Cold War Aerial Espionage*. Annapolis, MD: Naval Institute Press.

Byrne, Malcolm. 2014. *Iran-Contra: Reagan's Scandal and the Unchecked Abuse of Presidential Power*. Lawrence, KS: University Press of Kansas.

Cain, Bruce E. 2014. *Democracy More or Less*. Cambridge University Press.

Calabresi, Guido, Jose A. Cabranes and Gerald W. Heaney. 1998. "Kronisch v. United States (1998)." *United States Court of Appeals, Second Circuit* Docket No. 97-6116.

Carnegie, Allison. 2021. "Secrecy in International Relations and Foreign Policy." *Annual Review of Political Science* 24:213–233.

Carnegie, Allison and Austin Carson. 2018. "The Spotlight's Harsh Glare: Rethinking Publicity and International Order." *International Organization* 72:627–657.

Carson, Austin. 2018. *Secret wars : covert conflict in international politics*.

CIA. 1956. *Brainwashing from a Psychological Viewpoint*. Washington, D.C.: CIA CREST, CIA-RDP78-02646R000100100002-4.

CIA. 1958. *Project Corona*. Washington, D.C.: CIA CREST, CIA-RDP75B00514R000200050012-7.

CIA. 1960. *DDA-DDS History, 1953-1956, Chap V, Security Controls, 1953-1956*. Washington, D.C.: CIA Electronic Reading Room RDP72-00121A000100040001-9.

CIA, Inspection & Security Staff. 1950. *Special Research, Blueburd*. Washington, D.C.: CIA CREST, CIA-RDP83-01042R000800010003-1.

Coe, Andrew and Jane Vaynman. 2019. "Why Arms Control Is So Rare." *American Political Science Review* pp. 1–14.

Colaresi, Michael. 2012. "A Boom with Review: How Retrospective Oversight Increases the Foreign Policy Ability of Democracies." *American Journal of Political Science* 56:671–689.

Colaresi, Michael P. 2014. *Democracy declassified: the secrecy dilemma in liberal states*. Oxford University Press.

CREST. 2011. *(U) Project MK-ULTRA*. Intellipedia: CIA Electronic Reading Room 06760269.

Debs, Alexandre and Nuno P. Monteiro. 2014. "Known Unknowns: Power Shifts, Uncertainty, and War." *International Organization* 68:1–31.

Dienesch, Robert M. 2016. *Eyeing the Red Storm: Eisenhower and the First Attempt to Build a Spy Satellite*. Lincoln and London: University of Nebraska Press.

Downs, George W. and David M. Rocke. 1994. "Conflict, Agency, and Gambling for Resurrection: The Principal-Agent Problem Goes to War." *American Journal of Political Science* 38:362.

Drezner, Daniel W. 2019. "Technological change and international relations." *International Relations* 33:286–303.

Dulles, Allen W. 1953. *Project MKULTRA: Extremely Sensitive Research and Development Program*. Washington, D.C.: CIA Electronic Reading Room C06767515.

Early, Bryan R. and Erik Gartzke. 2021. "Spying from Space: Reconnaissance Satellites and Interstate Disputes." *Journal of Conflict Resolution* 65(9):1551–1575.

Earman, J.S. 1963. *Report of Inspection of MKULTRA*. Washington, D.C.: CIA Electronic Reading Room C06767515.

Eisenhardt, Kathleen M. 1989. "Agency Theory: An Assessment and Review." *The Academy of Management Review* 14:57.

Farrell, Henry and Abraham L Newman. 2019. "Weaponized Interdependence: How Global Economic Networks Shape State Coercion." *International Security* 44(1):42–79.

Fischer, Benjamin B. 2001. *The Central Intelligence Agency's Office of Technical Service, 1951-2001*. Washington, D.C.: Office of Technical Service.

Goldfien, Michael A. and Michael F. Joseph. 2023. "Perceptions of Leadership Importance: Evidence from the CIA's President's Daily Brief." *Security Studies* 32:205–238.

Goldfien, Michael, Michael Joseph and Daniel Krcmaric. 2023. "When do leader backgrounds matter? Evidence from the President's Daily Brief." *Conflict Management and Peace Science* .

Greer, Kenneth E. 1973. "Corona." *Studies in Intelligence* 17:1–37. https://www.cia.gov/static/3d24f7019bf7e718fd1d2a5c57e6a646/corona.pdf.

Grissom, Adam. 2006. "The future of military innovation studies." *Journal of strategic studies* 29(5):905–934.

Haines, Gerald K. 1998. "Looking for a Rogue Elephant: The Pike Committee Investigations and the CIA." *Studies in Intelligence* 42(5):81–92.

Hawkins, D G, D A Lake, D L Nielson and M J Tierney. 2006. *Delegation and Agency in International Organizations*. Cambridge University Press.

Hinton, Henry L. 2001. *Observations on GAO Access to Information on CIA Programs and Activities*. Washington, D.C.: Government Accountability Office.

Horowitz, Michael C. 2020. "Do Emerging Military Technologies Matter for International Politics?" *Annual Review of Political Science* 23:385–400.

Horowitz, Michael C. and Shira Pindyck. 2023. "What is a Military Innovation and Why It Matters." *Journal of Strategic Studies* 46(1):85–114.

Houghton, Vince. 2019. *Nuking the Moon: And Other Intelligence Schemes and Military Plots Left on the Drawing Board*. Penguin Books.

Houston, David J. 2000. "Public-Service Motivation: A Multivariate Test." *Journal of Public Administration Research and Theory* 10(4):713–728.

Ivan, Cristina, Irena Chiru and Ruben Arcos. 2021. "A Whole of Society Intelligence Approach: Critical Reassessment of the Tools and Means Used to Counter Information Warfare in the Digital Age." *Intelligence and National Security* 36(4):495–511.

Jacobsen, Anne M. 2015. *The Pentagon's brain: An uncensored history of DARPA, America's top secret military research agency*. New York: Back Bay Books.

Johnson, Loch K. 2022. *The Third Option: Covert Action and American Foreign Policy*. New York: Oxford University Press.

Jones, D. C., P. Kalmi and A. Kauhanen. 2006. "Human Resource Management Policies and Productivity: New Evidence from An Econometric Case Study." *Oxford Review of Economic Policy* 22:526–538.

Joseph, Michael F. 2021. "A Little Bit of Cheap Talk Is a Dangerous Thing: States Can Communicate Intentions Persuasively and Raise the Risk of War." *The Journal of Politics* 83:166–81.

Joseph, Michael F. 2023. "Do Different Coercive Strategies Help or Hurt Deterrence?" *International Studies Quarterly* 67.

Joseph, Michael F and Michael Poznansky. 2018. "Media technology, covert action, and the politics of exposure." *Journal of Peace Research* 55:320–335.

Joseph, Michael F., Michael Poznansky and William Spaniel. 2022. "Shooting the Messenger: The Challenge of National Security Whistleblowing." *The Journal of Politics* 84:846–860.

Jungdahl, Adam M and Julia M Macdonald. 2015. "Innovation inhibitors in war: Overcoming obstacles in the pursuit of military effectiveness." *Journal of Strategic Studies* 38(4):467–499.

King, Nigel. 1990. "Innovation at work: The research literature.".

Kinzer, Stephen. 2019. *Poisoner in chief: Sidney Gottlieb and the CIA search for mind control*. New York: Henry Holt and Company.

Kollars, Nina. 2017. "Genius and mastery in military innovation." *Survival* 59(2):125–138.

Kopel, Michael and Christian Riegler. 2006. "Delegation in an Ramp;D Game with Spillovers." *SSRN Electronic Journal* .

Kuo, Kendrick. 2020. "Military Innovation and Technological Determinism: British and US Ways of Carrier Warfare, 1919–1945." *Journal of Global Security Studies* .

Kuo, Kendrick. 2022. "Dangerous Changes: When Military Innovation Harms Combat Effectiveness." *International Security* 2(47):48–87.

Kurizaki, Shuhei. 2007. "Efficient Secrecy: Public versus Private Threats in Crisis Diplomacy." *American Political Science Review* 101:543–558.

Lai, Edwin L.-C., Raymond Riezman and Ping Wang. 2009. "Outsourcing of innovation." *Economic Theory* 38:485–515.

Laird, Burgess. 2020. "The Risks of Autonomous Weapons Systems for Crisis Stability and Conflict Escalation in Future U.S.-Russia Confrontations." *The RAND Blog* .

Land, Edwin H. 1954*a*. *194. Letter From Edwin H. Land, Chairman of the Technological Capabilities Panel of the Science Advisory Committee, Office of Defense Mobilization, to Director of Central Intelligence Dulles*. Washington, D.C.: Foreign Relations of the United States.

Land, Edwin H. 1954*b*. *A Unique Opportunity for Comprehensive Intelligence: A Summary*. Washington, D.C.: National Security Archive.

Laurie, Clayton D. 2001. *Congress and the National Reconnaissance Office*. Washington, D.C.: Office of the Historian, National Reconnaissance Office.

Laursen, K. and Nicolai J. Foss. 2003. "New human resource management practices, complementarities and the impact on innovation performance." *Cambridge Journal of Economics* 27:243–263.

Lee, Caitlin. 2019. "The Role of Culture in Military Innovation Studies: Lessons Learned from the US Air Force's Adoptino of the Predator Drone, 1993-1997." *Journal of Strategies Studies* pp. 1–35.

Lonardo, Livio Di, Jessica S. Sun and Scott A. Tyson. 2020. "Autocratic Stability in the Shadow of Foreign Threats." *American Political Science Review* 114:1247–1265.

Macdonald, Julia M. 2015. "Eisenhower's Scientists: Policy Entrepreneurs and the Test-Ban Debate 1954–1958." *Foreign Policy Analysis* 11:1–21.

Malis, Matt. 2021. "Conflict, Cooperation, and Delegated Diplomacy." *International Organization* 75:1018–1057.

Malis, Matthew. 2024. "Foreign Policy Appointments." *International Organization* .

Maret, Susan. 2018. "Murky Projects and Uneven Information Policies: A Case Study of the Psychological Strategy Board." *Secrecy and Society* 1(2):1–85.

Marks, John. 1979. *The search for the "Manchurian candidate": The CIA and mind control*. London: Allen Lane.

May, Ernest. 1998. Strategic Intelligence and U.S. Security: The Contributions of CORONA. In *Eye in the Sky: The Story of the Corona Spy Satellites*, edited by Dwayne A. Day, John M. Logsdon and Brian Latell. Smithsonian Institution Press.

McCroy, Alfred. 2006. *A Question of Torture: CIA Interrogation, From the Cold War to the War on Terror*. New York: Henry Holt and Company.

Merlin, Peter W. 2015. *Unlimited Horizons: Design and Development of the U-2*. Washington, D.C.: NASA Aeronautics Book Series.

Miller, Gary J. 2005. "THE POLITICAL EVOLUTION OF PRINCIPAL-AGENT MODELS." *Annual Review of Political Science* 8:203–225.

Miller, Nicholas L. 2022. "Learning to Predict Proliferation." *International Organization* 76:487–507.

NBC. 1977. *An Interview with Admiral Turner.* Washington, D.C.: CIA CREST, CIA-RDP99-00498R000200150007-0.

Neads, Alex, Theo Farrell and David J. Galbreath. 2023. "Evolving Towards Military Innovation: AI and the Australian Army." *Journal of Strategic Studies* .

Oder, Frederic E.E., James C. Fitzpatrick and Paul E. Worthman. 1988. *The Corona Story.* Washington, D.C.: Center for the Study of National Reconnaissance.

Pedlow, Gregory W. and Donald E. Welzenbach. 1992. *The Central Intelligence Agency and Overhead Reconnaissance: The U-2 and Oxcart Programs, 1954-1974.* Washington, D.C.: History Staff, Central Intelligence Agency.

Pocock, Chris. 2000. *The U-2 Spyplane: Toward the Unknown.* Atglen, PA: Schiffer Publishing.

Posen, Barry. 1984. *The sources of military doctrine: France, Britain, and Germany between the world wars.* Ithaca: Cornell University Press.

Poznansky, Michael. 2020. *In the shadow of international law : secrecy and regime change in the postwar world.*

Redacted. 1952. *Special Research, Blueburd.* Washington, D.C.: CIA CREST, 0000140401.

Richelson, Jeffrey T. 2002. *The wizards of langley: Inside the CIA's Directorate of Science and Technology.* Boulder, CO: Westview Press.

Rogers, Mike and Glenn Nye. 2019. "Why America Must Boldly Win the Technological Race Against China." *The Hill* Oct. 21.

Rosen, Stephen Peter. 1988. "New ways of war: understanding military innovation." *International security* 13(1):134–168.

Sechser, Todd S., Neil Narang and Caitlin Talmadge. 2019. "Emerging technologies and strategic stability in peacetime, crisis, and war." *Journal of Strategic Studies* 42:727–735.

Streatfield, Dominic. 2007. *Brainwash: The Secret History of Mind Control.* New York: St. Martin's Press.

Taylor, Mark Zachary. 2016. *The Politics of Innovation.* Oxford University Press.

Thomas, Gordon. 1989. *Journey Into Madness: The True Story of Secret CIA Mind Control and Medical Abuse.* New York: Bantam Books.

U.S. Senate. 1976. *Final report of the select committee to study governmental operations with respect to intelligence activities.* Washington, D.C.: U.S. Government Printing Office.

Vaynman, Jane. 2022. "Better Monitoring and Better Spying: The Effects of Emerging Technology on Cooperation.".

West, Michael A. and Neil R. Anderson. 1996. "Innovation in top management teams." *Journal of Applied Psychology* 81(6):680–693.

Wheelon, Albert. 1998. CORONA: A Triumph of American Technology. In *Eye in the Sky: The Story of the Corona Spy Satellites*, edited by Dwayne A. Day, John M. Logsdon and Brian Latell. Smithsonian Institution Press.

Wolford, S., D. Reiter and C. J. Carrubba. 2011. "Information, Commitment, and War." *Journal of Conflict Resolution* 55:556–579.

Zhang, Baobao, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz and Allan Dafoe. 2021. "Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers." *Journal of Artificial Intelligence Research* 71:591–666–591–666.

Zoghi, Cindy, Robert D. Mohr and Peter B. Meyer. 2010. "Workplace organization and innovation." *Canadian Journal of Economics/Revue canadienne d'économique* 43:622–639.

Appendix

Table of Contents

A Formal appendix

A.1 Characterizing the distributions

In the manuscript we focus on three quantities: e_0, e_1, λ. Here we provide more information about their properties, which follow from the definition of Bayes' Rule. We will use these properties in constructing some of the proofs.

Remark on $p(\pi)$. The prior is given as $p(\pi)$ with an expected value, $e_0 = \mathbb{E}[\pi] = \int \pi p(\pi) d\pi$.

Defining the posterior We notate the posterior distribution given an observed m, $p_1(\pi|m) \propto f_{m|\pi}(m|\pi)p(\pi)$, suppressing the proportionality constant. The post-message expected value of π is, $\mathbb{E}[\pi|m] \propto \int \pi p_1(\pi|m) d\pi$.

Remark Properties of posterior belief. $\partial \mathbb{E}[\pi|m]/\partial m > 0$, and that $e_0 > \mathbb{E}[\pi|m]$ if $m < e_0$, $e_0 < \mathbb{E}[\pi|m]$ if $m > e_0$, $e_0 = \mathbb{E}[\pi|m]$ if $m = e_0$.

Define m^* as the message that satisfies $\mathbb{E}[\pi|m = m^*] = c_D - \theta$. Note, that if m^* can be defined, it is unique, that every message $m > m^* \implies \mathbb{E}[\pi|m > m^*] > c_D - \theta$, and that m^* is increasing in c_D and decreasing in θ. Whether m^* can be defined depends on p, c_D, θ. If p is binomial, for example, m^* is not necessarily definable. But it can be defined in many important cases. For example, if p is normal, then m^* can always be defined.

Remark Properties of λ. If m^* exists, $\partial \lambda / \partial k_D < 0$.

Remark Properties of λ. If m^* exists, we can re-write $\lambda = \mathbb{E}[m \geq m^*]$. Then, $\partial m^*/\partial \lambda < 0$. If m^* does not exist, then $\lambda = 0$.

Remark Properties of e_1. e_1 is defined if m^* exists. If e_1 is defined, it must be a real-valued number that satisfies $e_1 > c_D - \theta$.

A.2 Lemma 3.1: When open institutions do not innovate

In the open institution, there are three strategy profiles that can lead to innovation. In the first R selects research, D does not approve research, then D selects innovation absent research. If this was on path, then D could not profitably deviate to rejecting innovation at the final decision-node. But D prefers to deviate to rejecting innovation if $e_0 - c_D < 0 \equiv e_0 < c_D$. This cannot be satisfied if condition 1 is.

In the second, R selects research, D approves research, then D approves innovation after research. Off the path, if D does not research D approves innovation. We cannot support the off-path action if condition 1 holds.

In the third, R selects research, D approves research, then D approves innovation after research. Off the path, if D does not research D does not innovate (condition 1 is satisfied). In this pathway, D approves innovation at the final on-path node if $\mathbb{E}[\pi|m]+\theta-c_D-k_D > -k_D \equiv \mathbb{E}[\pi|m] > c_D-\theta$. Note, we use this to define λ.

Working backwards, D's expected utility for authorizing research given expectations of on-path play is $\lambda(e_1+\theta-c_D)-k_D$. If instead, D does not research he gets 0. This solves for condition 2.

This completes the proof.

A.3 Proposition 3.2: Secrecy facilitates innovation (+ when it does not)

We re-state the equilibrium strategies. R's strategy is to select secret research. D's on-path strategy is to select secret research, and then approve innovation post-secret research if research yields a signal that shifts D's posterior belief $\mathbb{E}[\pi|m] > c_D-\theta$, and reject the innovation otherwise. Off the path, if R asks for permission to conduct research, D does not approve research and does not approve innovation.

Notice that if R does ask for permission, we are in a sub-game that exactly reflects the open institution. It follows that D's off-path strategy to reject research and reject innovation is supported if Conditions 1, 2 are. This yields a pay-off of $0, 0$. Similarly, if R rejects an idea at the first node, pay-offs are 0 for both players.

Turning to on-path actions, in the final node, D approve an innovation iff $\mathbb{E}[\pi|m]+\theta-c_D-k_Dx \geq -k_Dx$.

Working backwards, consider R's on-path decision to engage in secret research. R's value from secret research is: $(1-\lambda)0+\lambda(e_1+\theta-c_Rx)-k_R$. R prefers this to all her other options (which each yield 0) when equilibrium condition 3 is satisfied. It follows that R cannot profit from deviating to open research, or scrapping the project under the conditions stated in the equilibrium.

Later, we will need to know when we do not observe secret innovation in the secret institution.

Lemma A.1 *If conditions 2 is violated, then we cannot support secret innovation in the secret institution.*

When condition 2 is violated, we cannot support an SPE where D rejects open research. Thus, in every SPE D will approve open research if R selects open research. In this case, R's expected utility from requesting open research is $\lambda(e_1+\theta-c_Rx)-k_Rx$, which strictly dominates R's value from secret research.

Lemma A.2 *If condition 2 is violated and condition 1 is not, and*

$$\lambda > \frac{k_rx}{e_1+\theta-c_Rx} \tag{6}$$

then then we can support open research in the secret institution. In equilibrium, D approves open research, does not approve innovation without research, and approves innovation following research (secret or open) iff $\mathbb{E}[\pi|m] > c_D-\theta$. R selects open research.

We've already shown that: (i) when condition 1 is violated, D rejects innovation absent research; and that (ii) when condition 2 is violated, D prefers to approve open research than reject it given (i); and that (iii) R strictly prefers open research given to secret research given (ii). Finally, R prefers open research to scrapping the project if $\lambda(e_1 + \theta - c_R x) - k_R x > 0$, which solves for the equilibrium condition. This completes the proof.

A.3.1 Existence of Proposition 3.2

We have solved for a set of conditions where Proposition 3.2 is incentive compatible. We now verify that these conditions can be satisfied for some parameters. To be clear, these are not exhaustive conditions. Our only goal is to demonstrate that the equilibrium conditions can be satisfied.

Remark Assume a prior distribution $\pi \sim \mathcal{N}(0, \sigma_0)$, and parameters $\theta > c_D > c_R x$. Then, there must exist a $k_R, k_D > 0$ for which equilibrium conditions (1), (2) and (3) are simultaneously satisfied.

First note that $e_0 = 0 \implies e_0 < k_D$. Thus, condition (1) is satisfied.

We now prove that m^* must exist given the assumed parameters and distributions and discuss its implications for e_1, λ. Note the posterior distribution is

$$\mathcal{N}(\frac{\sigma m}{\sigma + \sigma_0}, \frac{1}{\sigma + \sigma_0})$$

with an expected value of $\frac{\sigma m}{\sigma + \sigma_0}$. Note because the domain of both prior and posterior cover all real valued numbers, we can always find an m that satisfies:

$$\frac{\sigma m}{\sigma + \sigma_0} = c_D - \theta \implies \exists\ m^*, m^* := \frac{(c_D - \theta)(\sigma + \sigma_0)}{\sigma}$$

This implies λ and e_1 exists and λ satisfies > 0. By construction, if e_1 exists, then $e_1 + \theta - c_D > 0 \implies e_1 + \theta - c_R x > 0$.

We now turn to conditions (2) and (3), which we re-write as:

$$\lambda(e_1 + \theta - c_D) < k_D$$

$$k_R < \lambda(e_1 + \theta - c_R x)$$

The only restriction on k_D, k_R is that they must be positive. We can trivially find a k_D sufficiently large to satisfy condition (2). To satisfy (3) a k_R must exist that satisfies $k_R \in (0, \lambda(e_1 + \theta - c_R x)$. We've shown that $\lambda > 0, e_1 + \theta - c_R x > 0 \implies \lambda(e_1 + \theta - c_R x) > 0$. Thus, the open interval $(0, \lambda(e_1 + \theta - c_R x)$ is always defined under the stated conditions.

A.4 Two pathways to innovation

In this section we explain how we derive our empirical implications from the main model. The basic idea is to conjecture a set of parameters where secrecy facilitates innovation (i.e, conditions 1, 2, and 3 all hold), then show how taking certain parameters to their limits implies we must violate either conditions 1, 2, and 3. If we violate them, then secrecy cannot faciliate innovation.

The first pathway describes different features of $p()$. Our first bullet point relates to e_0, the expected value of p. Define a second distribution p_α as identical to p in functional form, but with a shifted mean $\alpha \in \mathcal{R}$. That is, for an arbitrary input a, $p(a) = p_\alpha(a - \alpha)$. Note the prior expected value of $p_\alpha = e_0 + \alpha$.

Define the standard deviation of $p()$ as σ_p, not that it equals the standard deviation of p_α.

We can re-write the first bullet point in pathway 1 as follows. Suppose $\alpha = 0$, and otherwise we take a list of parameters that meet the conditions for secret innovation outlined in proposition 3.2. The claim in the bullet is that we will violate at least one equilibrium condition if $\alpha \to \infty, -\infty$.

Starting with the upper bound, there exists a $\bar{\alpha}$ large enough so that for any $\alpha > \bar{\alpha}$ inequality of condition 1 is violated. Since it is violated, we cannot say that innovation does not occur in the open institution, and therefore we cannot say that secrecy facilitates innovation.

Turning to the lower bond, there exists a $\underline{\alpha}$ low enough so that for any $\alpha < \underline{\alpha}$ inequality of condition 3 is violated. Both λ, e_1 are a function of α. By Bayes' Rule, as $\alpha \to -\infty$, $\lambda \to 0$. As stated above, as $\alpha \to -\infty$ e_1 is either undefined (which assures condition 3 cannot be satisfied) or must satisfy $e_1 > c_D - \theta$, as desired.

Our second bullet point relates to the standard error of p, σ_p. Define a second distribution p_β as p with a resolution parameter $\beta > 0$. More precisely, for an arbitrary input a, $p(e_0 + a) = p_\beta(e_0 + \beta a)$. Notice that the expected value of p_β is equal to e_0. We also define the standard error as σ_β. For $\beta < 1, \sigma_\beta < \sigma$, and for $\beta > 1, \sigma_\beta > \sigma$.

We can re-write the second bullet point as follows. Suppose a list of parameters where condition 1,2 and 3 are satisfied and replace p with p_β, $\beta = 1$. There exists a $\underline{\beta}$ small enough so that for any $\beta < \underline{\beta}$ that violates condition 3.

Taking $\beta \to 0 \implies \mathbb{E}[\mathbb{E}[\pi|m]|p_\beta] \to e_0$. Thus, taking $\beta \to 0$, if $e_0 < c_D - \theta \implies \lambda \to 0$ and e_1 is undefined. Note pathway 2 caveats that as θ cannot be too large, and we can more precisely characterize that as the condition $\theta < c_D - e_0$. Substantively, this means that the improvement value of research is insufficient to induce D to approve a program. This is consistent with the overall message of the argument, wherein research reveals information.

The second pathway simply highlights that proposition 3.2 can hold when research is costly and the idea is promising. It starts with the premise that managers know a project shows promise once it is improved by research ($e_0 + \theta >> 0$). However, they know that the research involves political costs that outweigh the amount that research improves the project (i.e. $\theta \leq k_D$).

We start by noting that that for any initial expectation of success e_0, and amount that research improvement θ, there exists a sensitivity to the costs of authorization c_D for which condition 2 holds. Similarly, for any e_0, there exists a sensitivity to the costs of research $k_D > e_0$ for which condition 1 holds.

Now focusing only on conditions where $e_0 > 0$ and condition 2 holds, and where m^* exists. Then there exists a $k_R \to 0$ for which we can satisfy 3. The proof follows instantly from the proof of proposition 3.2 (especially noting that 3 is always satisfied if $k_R = 0$.). So long as there is some chance that research will convince D, we can find a researcher sufficiently insensitive to costs who is willing to research given that small chance.

A.5 Expectation 1 and 2 (section 3.2.2)

Expectation 1 is validated if the λ values that support open research are larger than the λ values that support secret research. Expectation 1 follows from three facts. (1) Secret research cannot occur if condition 2 holds, which places an upper bound on λ. (2) Open research only requires that λ is sufficiently large (lemma A.2). (3) $\lambda = Pr.$(Research Approved).

Turning to expectation 2. To provide some intuition, Expectation 2 takes the perspective of an outside observer who knows the true value of π. Holding the value of π constant at an effect that is sufficiently effective (which we shall define below), we ask the outside observer to considers two worlds: one in which secret research appears on path (call it the counter-factual case), and another where open research appears on path (call it the baseline case). In both cases innovation can occur with probability, but only if the message is sufficiently strong. How confident is the outside observer that innovation will actually occur in each case given that π is known? Then, if we increase π even more, how does it affect the outside observer's confidence that innovation will occur in the baseline relative to the counterfactual? The basic idea is that the minimum message m that will induce innovation in the counterfactual world must be stronger than the minimum message necessary to induce innovation in the baseline. Given that $m \sim \mathcal{N}(\pi, \sigma)$, increasing π has a greater impact on the observer's expectation of innovation in the counter-factual case.

To make this claim more precise way, we first must narrow our focus to two comparable cases. One that leads to secret research on path (call it the counter-factual case) and the other that leads to open research on path (call it the baseline case). To do it, we again utilize distribution that are identical in their function forms, but vary in their expectations. We call p our baseline distribution. We assume it is supported positively on $\mathbb{R}$. We then define a counter-factual distribution p_α as identical to p in functional form, but with a shifted mean $\alpha < 0$. That is, for an arbitrary input a, $p(a) = p_\alpha(a - \alpha)$. To differentiate between cases, we index expectation derived from p_α as $\lambda_\alpha, e_{0\alpha}, e_{1\alpha}$. Given $\alpha < 0$, note that $e_{0,\alpha} < e_0, e_{1,\alpha} < e_1, \lambda_\alpha < \lambda$.

We assume fixed values of prior parameters $k_i, c_i, x, \theta, \sigma$, and the functional form of p. We then assume that given the baseline case of p we satisfy the conditions for open research characterized in lemma A.2, and that given the counterfactual case of p_α we can we satisfy the conditions for secret research characterized in proposition 3.2.

In both cases, research appears on path and there is a positive probability that D approves innovation occurs post-research. We've shown that in either case, D approves innovation post-research iff: $\mathbb{E}[\pi|m, p] > c_D - \theta$, where m is a function of π. Define $m^\dagger$ as the message necessary to satisfy $\mathbb{E}[\pi|m^\dagger, p] = c_D - \theta$, and $m^\dagger_\alpha$ as the message necessary to satisfy $\mathbb{E}[\pi|m^\dagger_\alpha, p_\alpha] = c_D - \theta$. By construction of the counter-factual case, $m^\dagger_\alpha > m^\dagger$.

Suppose an outside observer knows the true $\pi = \pi_x$. That outside observer's pre-research expectation that innovation occurs in our two worlds is characterized as:

$$\omega_x = \int_{m^\dagger}^{\infty} \mathcal{N}(\pi_x, \sigma) dm$$

$$\omega_{x,\alpha} = \int_{m_\alpha^\dagger}^{\infty} \mathcal{N}(\pi_x, \sigma) dm$$

These are the beliefs that a message $m > m^\dagger$ or $m > m_\alpha^\dagger$ will occur that will give D enough confidence to approve innovation. We similarly $\omega_y, \omega_{y\alpha}$ for $\pi_y > \pi_x$.

We can re-state expectation 2 as follows: Contrasting our two worlds, so long as the true effect of innovation is sufficiently large ($\pi_x > m^\dagger$), then increasing the true effect of innovation from $\pi = \pi_x \to \pi_y$ raises the probability of innovation in the counterfactual world more than the baseline world:

$$\omega_y - \omega_x < \omega_{y\alpha} - \omega_{x\alpha}$$

Re-arranging this term, the claim is true if:

$$\int_{m^\dagger}^{m_\alpha^\dagger} \mathcal{N}(\pi_y, \sigma) dm > \int_{m^\dagger}^{m_\alpha^\dagger} \mathcal{N}(\pi_x, \sigma) dm$$

true if $\pi_x > m^\dagger$. As desired.

A.6 Proposition 3.3: Monitoring and principal-agent dynamics

To start, we re-state the timing and information of the model more precisely.

Initially, Nature draws two random variables:

- $\pi \sim p()$ (unobserved by R or D)
- $k \sim f()$ (privately observed by R).

Then the first decision node is R's, where R decides between open research, secret research, or scrapping the project.

If R pursues open research:

- D observes R's request, and the value of k,
- the game proceeds as in baseline starting at the node where D can approve open research/not.

If R chooses to scrap the project,

- D observes that R has not conducted open research, but not k.

- D can monitor R or not. If D monitors, R's choice (scrap project) and k is observed by D and not observed otherwise.
- Regardless of D's choice to monitor the game ends with payoffs $0, 0$.

If R chooses secret research,

- D observes that R has not conducted open research, but not k.
- D can monitor R or not. If D does not monitor,
  - Neither R's choice (secret research) nor k are observed.
  - the game proceeds with secret research as in the baseline model with Nature's draw $m|\pi$. All payoffs are identical to the baseline.
- If D monitors:
  - D observes both R's choice of secret research, and k. D can chose to shut down research or not.
  - If D shuts down research
    - R incurs k, D incurs no research cost.
    - No player observes m.
    - D is given the choice to approve innovation or not, given $\mathbb{E}[\pi] = e_0$.
  - If D does not shut down research.
    - R incurs xk, D incurs k
    - Both players observe m
    - D is given the choice to approve innovation or not, with $\mathbb{E}[\pi|m]$.

Once research has occurred (or if it does not), the incentives for choices are equivalent. Thus, D will not approve innovation absent research if condition 1 holds. D will only approve innovation post research if $\mathbb{E}[\pi|m] > \theta - c_D$.

We now turn to D's decision to approve research. There are two ways D has the option to approve research. First, R may select open research. In this case, D will only approve if

$$\lambda[e_1 + \theta - c_D] > k \tag{7}$$

This is a re-statement of condition 2 subbing $k_D = k$. Second, R selects secret research and D monitors. In this case, D also only approves research if 7 holds.

We now conjecture that D will not monitor and identify R's incentive to select secret research, open research, or scrap the project for different values of k, assuming the conjecture holds. In a moment, we will consider D's incentive to monitor. We've shown if condition 7 is violated, then R's expected value from asking for open research is 0. R's value for scrapping a project is also 0. Thus for a k that violates condition 7, we could support either choice if 0 is better than R's expected utility from secret research. We've show above that if D will approve open research (7 holds), then R strictly prefers open research to secret research. Given these results, we need only consider when

R prefers secret research to a payoff of 0. R prefers secret research to if $\lambda[e_1 + \theta - c_R x] > k$. We assumed, $0 < \lambda[e_1 + \theta - c_D] < \lambda[e_1 + \theta - c_R x]$, which imposes an order on these conditions.

Putting it altogether, if D will not monitor, we can support R's on-path strategy, defined by two cut points on k. Let $\underline{k} = \lambda[e_1 + \theta - c_D]$. If $k < \underline{k}$, R selects open research, and D approves, as desired. Let $\bar{k} = \lambda[e_1 + \theta - c_R x]$. If $k > \bar{k}$, D will not approve open research if asked, R is indifferent between asking and being rejected and open research, and both these options are better than secret research. In equilibrium, we conjecture that R scraps the idea. If $k \in [\underline{k}, \bar{k}]$, R pursues secret research.

We now resolve our conjecture that D will not monitor in the case that D has not observed research. Off path, if D monitors after failing to observe open research, D gets 0 for any potential value of k. If $k \in [\underline{k}, \bar{k}]$, D rejects, leading to 0. If $k > \bar{k}$, R did not research and the idea is scrapped. D's expected utility from on path play (not monitoring) is:

$$pr[k > \bar{k}|nor] \times 0 + pr[k \in [\underline{k}, \bar{k}]|nor](\lambda(e_1 + \theta - c_D) - x\mathbb{E}[k|sr, nor])$$

Here $pr[k \in [\underline{k}, \bar{k}]|nor]$ is D's expectation R undertook secret research given that D did not observer research (nor represents no research observed). Then $\mathbb{E}[k|sr]$ is D's expected value of k given the values of k for which R conducted secret research under the condition that no research was observed (sr represents secret research occured). D prefers not monitoring to monitoring given D did not observe research if:

$$\frac{\lambda(e_1 + \theta - c_D)}{x} > \mathbb{E}[k|sr, nor] \tag{8}$$

as stated in the equilibrium.

A.7 Lemma 3.4: Researcher can fabricate her report

First we more fully specify the set-up of the extension with some reference to the baseline presented in Figure 1.

- D can costlessly set k_R, c_R (which represents a manager hiring a particular researcher).
- $\pi \sim p()$ (unobserved by R or D)
- R selects between open research, secret research, or scrapping the project.
  - Open research and scrapping the project proceed identically as in the baseline presented in Figure 1.
- If R selects secret research, R privately observes $m \sim \mathcal{N}(\pi, \sigma)$
- R writes costless message $m_R \in \mathcal{R}$, which is public.
- D decides to innovate or not.

The payoffs in this extension are identical to the baseline model, with the subtle difference that k_R, c_R are endogenous to D's choice.

The model includes a new information structure such that R has the same information as in the baseline model. But R also has private information about the message m in the case of secret research.

The new information structure means we must provide additional information about beliefs. We continue to define $e_0 = \mathbb{E}[\pi|p]$, $\lambda = pr(\mathbb{E}[\pi|m] > c_D - \theta)$, $e_1 = \mathbb{E}[\mathbb{E}[\pi|m]|\mathbb{E}[\pi|m] > c_D - \theta]$.

These expectations will play the same role in the analysis, in the event of open research, and for R's choice to engage in secret research. However, beliefs could deviate following secret research. We define $e_R = \mathbb{E}[\pi|m]$, as R's expected value of π given R has observed research. We define, $e_D = \mathbb{E}[\pi|m_R, s^R]$ as D's expected value of π given D observes R's message and R's strategy.

We define an honest researcher as one who sends the message $m_R = m \forall m$. We define a trust-worthy researcher as one that induces $e_D = e_R|m_R, s^R$ for all possible m. Meaning that D's beliefs match R's beliefs at the moment D must chose to approve innovation or not.

Lemma A.3 *If conditions 1-3 hold, for $c_R = c_D/x$, and*

$$\lambda > \frac{k_D x}{e_1 + \theta - c_D} \tag{9}$$

then the following strategies are supported as a PBE.

D sets $k_R < \lambda(e_1 + \theta - c_R x), c_R = c_D/x$, then

- *D approves innovation following secret research iff $m_R \geq m^*$. Off-path, D rejects open research, and rejects innovation absent research.*
- *R selects secret research, and sets $m_R = m$.*

If, off-path, D sets $k_R > \lambda(e_1 + \theta - c_R x)$, or $c_R \neq c_D/x$, then we revert to the following off-path strategies:

- *R selects scrap the project, and sends a message $m_R \sim p(\pi)$ which is not conditioned on then observed m, and covers all feasible messages with positive probability (i.e, no off-path messages).*
- *D rejects research and innovation at every decision-node.*

Remark In equilibrium, R is honest $m_R = m$ and trustworthy ($e_R = e_D$).

We claim that if D deviates by setting the incorrect k_R, c_R, then R scraps the idea at the first decision node, leading to a payoff of 0 for both players. This is supported by a series of other off-path actions. We now solve for this off-path profile. Remaining in the case where D deviates to setting an off-path cost profile, we conjecture that if R did pursued secret research given an incorrect cost profile, that R also sends a babbling message that covers all feasible messages, and

D rejects innovation. Note because R's message is babbling, $e_D = e_0$. Thus, D rejects innovation if $e_0 < \theta - c_D$, true if conditions 1, 2 hold. If D will not approve innovation for any m, R strictly prefers to scrap the idea over secret innovation, as desired. We note that no matter R's cost profile, D rejects open research if condition 1 - 3 hold. Thus, R also cannot profitably deviate to open research, as desired. It follows that if D sets the incorrect cost profile, that we can support a strategy profile where players expect 0.

We now turn to the on-path case where D sets $k_R < \lambda(e_1 + \theta - c_R x), c_R = c_D/x$. I claim that $c_R = c_D/x$ induces R to truthfully reveal information. We can support R's truthful revelation if R's incentives for accepting over rejecting innovation are identical to D's for any m. This is true if $c_R = c_D/x$. I claim that D wants to induce R to conduct secret research. Here the relevant counter-factual is that D selects some other researcher, leading to a payoff of 0. D prefers to induce secret research if $\lambda[e_1 + \theta - c_D] - k_D x > 0$, which solves for condition 5. Finally, a cost profile must exist so that R wants to select secret research rather than deviate to either open research or scrapping the project (both yield expected value of 0). True iff, $\lambda[e_1 + \theta - x c_R] - k_R > 0$, as desired.

A.8 External ambiguity, and calibrating cost passing

We start with the baseline model. We then adjust it at only one decision node. If R asks for open research, that ask is continuous and thus there are many forms of open research. Specifically, at the first decision node (R selects between scraps, secret or open research), if R does not scrap the project, R's choice to research is represented by a continuous variable z. We allow R to set $z \in [0, 1 - x]$. If R sets $z = 0$, the model goes down the secret innovation pathway exactly as in the baseline model. If R sets $z > 0$, the model goes down the open research pathway with choices exactly as in the baseline model. However, we assume that the cost share parameter in the payoffs is adjusted, so that D accrues a $x + z$ share of the research cost if D approves research, and R accrues a $1 - z$ share of the research cost if D approves research.

Specifically, we only see a payoff adjustment under two conditions. First, if R asks for open research, D approves open research, and D rejects innovation, payoffs are: $U^D = -k_D(x+z), U^R = -k_R(1-z)$. Second, if R asks for open research, D approves open research, and D then approves innovation, payoffs are, $U^D = \pi + \theta - k_D(x+z) - c_D, U^R = \pi + \theta - k_R(1-z) - c_R x$.

Notice that $z = 1 - x$ is equivalent to the baseline payoffs from open research. But when $z < 1 - x$ R takes on a larger share of the burden from open research. If $z = 0$, the payoffs are the same as in secret research. Loosely, we can think about z as representing the expected chance that agents within the national security agency can keep the manager's knowledge of devilish details secret from some un-modeled, higher-level principal. This represents a case where D knows what R is doing, but R has informed D in such a way, that higher level principles may assign the blame to R. See the manuscript for more substantive motivation.

This variant of the model represents a tough theoretical test for the relevance of internal secrecy because only $x = 0$ represents true internal secrecy. We assume that the researcher is going to the manager for all $x > 0$, the manager fully understands what the research involves and can shut down a project if he wants to. We'll show that even under this tough test, conditions arise when the researcher still exploits internal secrecy.[33]

[33] We get even stronger results in favor of internal secrecy in a model where increasing z both increases the manager's cost, and probabilistic informs the manager of the devilish details. This would represent a setting where

First, we solve for the case where we observe cost sharing and partially external secret but internally open research.

Define $z^* = \min[1 - x, \frac{\lambda(e_1+\theta-c_D)}{k_D} - x]$.

Proposition A.4 *If condition 2, 1 holds, and*

$$\frac{\lambda(e_1 + \theta - c_D)}{k_D} - x > z > 1 - \frac{\lambda(e_1 + \theta - c_R)}{k_R}$$

can be jointly solved for some $z \in (0, 1 - x]$, then the following strategies are SPE. R sets $z = z^*$ in the research phase. D's strategy is to accept research if $z \leq z^*$ and deny research otherwise. Regardless of how research occurs, D approves innovation if $\mathbb{E}[\pi|m] \geq c_D - \theta$, and reject innovation otherwise. Off path D rejects innovation absent research.

The extension does not adjust payoffs for innovation. Thus, as shown conditions 2, 1, guarantee that D will strictly reject innovation absent research, and reject innovation post-research if $\mathbb{E}[\pi|m] < c_D - \theta$ and accept it otherwise.

Turning to the choice to research. D does not make a choice if R selects secret research. If R selects a variant of open research, D approves research if: $\lambda(e_1 + \theta - c_D) - k_D(x + z) > 0$. This solves for the LHS of the equilibrium condition. Thus, when this condition is satisfied, D cannot profitably deviate to rejecting open research.

If R sets z too high, or scraps the idea, R's expected payoff is 0. R prefers to set z at some level D will accept over a choice that induces a payoff of 0 iff $\lambda(e_1 + \theta - c_R) - k_R(1 - z) > 0$. This solves for the RHS of the equilibrium.

We claim that if the equilibrium condition is satisfied, R sets z^*. z^* defines the the largest amount of cost-sharing that is both feasible (By assumption, bounded at $z \leq 1 - x$) and that D will accept. We've shown that R cannot profitably deviate to a higher z^* under stated conditions because D will reject. Since R' utility is increasing in D's responsibility, R cannot profitably deviate to sets $z < z^*$. This completes the proof.

As expected, R's incentives are to defray the political costs of research by passing them onto R's manager. This creates incentives for R to pursue open research over secret research $z = z^* > 0$. One might wonder, do researchers ever sustain internal secrecy from their managers if they have the option to pass on costs? We now identify the conditions where we still observe secret research,

Proposition A.5 *If conditions 1, 2, 3, and*

$$\frac{\lambda(e_1 + \theta - c_D)}{k_D} < x \tag{10}$$

hold, then the following strategies are sub-game perfect. R selects secret research: $z = 0$. D's strategy is to deny all requests for research, deny innovation absent research, and approve innovation post-research if $\mathbb{E}[\pi|m] \geq c_D - \theta$ but not otherwise.

the research writes a vague report, or a very technical report where the devilish details are buried. In a situation like this, the manager may pick up the details but may not.

D will reject every open research request $z \in (0, 1-x]$, if $\lambda(e_1 + \theta - c_D) - k_D x < 0$. This gives us equilibrium condition 10. Thus, if R selects open research R's expected utility is 0. Note we are now in an identical situation to the baseline model. The result from proposition 3.2 carries through. We can support secret innovation if conditions 1, 2, 3 hold, as desired. Note condition 10 is easier to satisfy when c_D, k_D are high. This substantiates our claim in the manuscript that researchers only exploit informal briefs when the manager's costs are low, and that we expect to see this kind of informal briefing in the deep uncertainty pathway. However, we still expect true internal secrecy over the devilish details when condition 10 is satisfied.

A.9 Institutional design

We now introduce a higher-level principal who: (a) has a stake in the national security welfare of the country; (b) has the power to write the rules that govern how members of the executive incur costs. In the U.S. context, this principal could represent Congress.

We start with the monitoring extension presented in section A.6. In terms of timing, we add but one choice to the beginning of the game. We allow Congress to set $x \in [0, 1]$. All agents publicly observe x. At that moment, Congress becomes passive, and the game unfolds between R and D given the set x as it is presented in section A.6. Note, this framework closely matches how Congress writes rules for the national security community. Specifically, Congress pass general laws that determine the conditions under which a specific agent will face costs. These include laws that determine what actions are illegal, or constitute professional misconduct. It also includes who has a responsibility for their subordinates, and who has a responsibility to speak up if their managers abuse the law. Members of the intelligence community are then confronted with specific scenarios (e.g. the decision to pursue a particular idea) knowing what the laws that govern their actions are, the risks of exposure, etc.

As we shall see, setting x has two affects. First, it alters the strategic incentives of the agents in the research institution. Second, it imposes a direct cost on Congress because, consistent with our motivation that internal secrecy is important to sustain external secrecy, it raises the risk that foreign rivals will discover the programs and capabilities of our national security institutions.

We assume that Congress' utility function is similar to the manager's in that Congress incurs the research and innovation costs when the manager does. We assign c_O (O for overlord) as Congress's cost for pursuing innovation. We assume Congress suffers the common k, which is randomly drawn in this model and discussed in section A.6.[34] We allow the possibility that Congress suffers one additional cost, $g(x)$, which is weakly increasing in x and $g(0) = 0$. This cost represents the inevitable trade off between internal secrecy and external secrecy. As discussed in the concepts section, internal secrecy is what partly excuses agents from punishment when their team makes choices that they did not know about, or had limited ability to question. In an open institution $x = 1$, meaning that all agents are responsible for finding out what is happening in their own team and reporting wrongdoing when they see it. But as discussed in the concepts section, the higher x is, the greater risk there is that foreign rival will discover our intelligence practices. Putting these pieces together, Congress' utilities are:

[34] Note that since Congress takes the first action, Nature has not yet drawn k when Congress acts. It does not matter of Congress observes k or not, because Congress has no additional actions.

$$U^O(\text{research, innovation}) = \pi + \theta - k - c_o - g(x)$$

$$U^O(\text{no research, innovation}) = \pi - c_o - g(x)$$

$$U^O(\text{research, no innovation}) = -k - g(x)$$

$$U^O(\text{no research, no innovation}) = -g(x)$$

The theoretical concern that motivates this extension is as follows. Even if it is true that a high amount of internal secrecy would incentivize agents to participate in the don't-ask-don't-tell scenario, Congress would anticipate this concern and change the institutional rules so that national security agents would not exploit it. Thus, our goal is to show that conditions exist where Congress would prefer to live with don't-ask-don't-tell, rather than prevent it.

We proceed as follows. First, we focus our analysis on conditions where we can induce don't-ask-don't tell for $x < x^*$, but Congress can prevent this behavior by setting $x \geq x^*$. Second, we isolate the two conditions—$x < x^*$, $x \geq x^*$—and separately solve for strategy profiles for R and D that we can support on path in each case. Along the way we identify Congress' utility from setting x in either range, given R and D play these strategies. In the $x \geq x^*$ case, we show that Congress induces D to monitor if ever D does not observe research. This creates the following test for our analysis. If O ever sets $x < x^*$ in equilibrium, then we can say that O is not willing to set x sufficiently large to prevent agency loss. Third, we characterize an equilibrium. Finally, we solve for a minimum condition where Congress has a profitable deviation from every $x \geq x^*$ to $x = 0$, given the strategies for R and D that x will produce. Thus, we identify conditions where Congress would not set x large enough to scuttle don't-ask-don't-tell because Congress has at least one profitable deviation to $x = 0$.

A.9.1 Parameter restrictions

To start, we focus on fixed set of values that allow us to support the don't-ask-don't-tell equilibrium defined in proposition 3.3 for a range of x. Using the same definition for $\mathbb{E}[k|sr, nor]$ as above, define $x^* = \frac{\lambda[e_1+\theta-c_D]}{\mathbb{E}[k|sr,nor]}$. This is the value of x for which condition 4 becomes an equality.

Then, define a set of fixed values of, $p(), \sigma, f(k), c_R < c_D, \theta$ for which we can support the don't-ask-don't-tell equilibrium defined in proposition 3.3 for all $x \in [0, x^*]$. Note that x appears in conditions 3 4, and both are easier to satisfy as x decreases. In the limit, at $x = 0$, condition 8 is always satisfied, and condition 3 reduces to $k < \lambda(e_1 + \theta)$. Finally, $\bar{k} \to \lambda(e_1 + \theta)$.[35] Also note that conditions 1 and 2 do not depend on x, and are assumed satisfied by our parameter restriction.

Summing up the implications of these restrictions for R and D's strategy. By construction, if $x \geq x^*$ we cannot support proposition 3.3 because condition 4 is violated. But if $x < x^*$ condition

[35] While it is true that adjusting x influences $\bar{k}$ which increases $\mathbb{E}[k|sr]$. It is also the case that $\mathbb{E}[k|sr]$ is a real valued number for any x, thus, if $x = 0$, inequality 4 is always satisfied.

4 holds, and we can support no monitoring and secret innovation. As a result, Congress can induce don't-ask-don't tell if Congress sets $x < x^*$, but will prevent it otherwise.

In what follows we separately analyze the $x < x^*, x \geq x^*$ cases. Where not specified, we define strategies for R and D we can support in a PBE given x that falls into these respective ranges. We also specify O's expected utility given those strategies.

A.9.2 The $x < x^*$ case

By construction, we can support the strategy for R and D as stated in proposition 3.3.

Remark Conjecture that if Congress sets $x < x^*$ that R and D play the strategies described in proposition 3.3. Then, Congress's expected utility for setting $x < x^*$ is:

$$EU^O(x < x^*) = pr[k < \lambda(e_1 + \theta - c_R x)]](\lambda(e_1 + \theta - c_O) - \mathbb{E}[k|k < \lambda(e_1 + \theta - c_R x)]) - g(x)$$

Note we cannot make strong claims about which $x \in [0, x^*)$ maximizes Congress' utility because Congress faces a three-way trade-off between increasing the direct cost, $g(x)$, increasing the probability that R will pursue research, which increases the value that Congress expects to accrue because more profitable programs get funded, and increasing the expected cost that Congress incurs should research happen. It is possible that there are multiple maximum values, and they may take on the corner $x = 0$.

A.9.3 The $x \geq x^*$ case

We now characterize one strategy profile for R and D that we can support on path, under the assumption that $x \geq x^*$.

Lemma A.6 *Fixing Congress's strategy at $x \geq x^*$, then we can support the following strategies on path in a PBE.*

- *D always monitors if D fails to observe open research. Regardless of how D comes to identify research, D approves research if $k \leq \lambda[e_1 + \theta - c_D]$ and rejects it otherwise. D rejects all innovation absent research, and approves innovation post-research iff $\mathbb{E}[\pi|m] > c_D - \theta$.*
- *R requests open research if $k \leq \lambda[e_1 + \theta - c_D]$ and scraps the project otherwise.*

Because condition 1 and 2 are satisfied and constant for any x, we know that D will reject innovation absent research, D will approve innovation iff $\mathbb{E}[\pi|m] > \theta - c_D$, and D will reject open research if $k > \lambda[e_1 + \theta - c_D]$. What is more, we can still use the same definitions for e_0, e_1, λ, which do not depend on x.

We conjecture that if $k > \lambda[e_1 + \theta - c_D]$, R always scraps the project, and D always monitors. Further, if after monitoring, D did observed secret research, D would reject research. At the moment R decides to scrap the project, R's expected utility from on path play is 0. At the moment, D decided to monitor, D's expect utility from on-path play is 0.

Starting with D's incentive to reject secret research if discovered. As shown in proposition 3.3, D's reject research post-monitoring if $k > \lambda[e_1 + \theta - c_D]$. Thus, D cannot deviate from rejecting research if D monitors and discovers that R has conducted secret research. Turning to D's incentive to monitor. Note R does not play secret research on path. Thus, if D does not observe research, D expects 0 from not monitoring. Thus, D is indifferent between monitoring (on path) or not. Turning to R's incentive to scrap the project. As shown in proposition 3.3, if R deviates to open research, D rejects open research (and then innovation) if $k > \lambda[e_1 + \theta - c_D]$. This leaves R indifferent between scrapping the project and selecting open research. If R selects secret research, R expects $-k$ given that D always monitors and shots down research. Clearly, R does worse from deviating to secret research.

We conjecture that if $k < \lambda[e_1 + \theta - c_D]$, R always requests open research and D approves. If R deviated to secret research D would monitor and approve. Thus, R's expected utility from on path play, at the moment R requests open research is $\lambda[e_1 + \theta - c_R x] - kx$, D's expected utility at the moment D approves open research is: $\lambda[e_1 + \theta - c_D] - k$.

As shown in proposition 3.3, D prefers to accept open research to not if $k < \lambda[e_1 + \theta - c_D]$. If R deviates to secret research, D observes no research. As just shown, D always monitors given this observation. But in this off-path case, D's monitoring discovers secret research and $k < \lambda[e_1+\theta-c_D]$. As just shown, D would approve. Thus, R is indifferent between secret and open research. Finally, consider R's deviation to no research. In this case, R gets 0. R can only profitably deviate to scrapping the idea if, $k < \frac{\lambda[e_1+\theta-c_R x]}{x}$. This is always satisfied if $c_R < c_D$ (true by the construction of the scenario) and also $k < \lambda[e_1 + \theta - c_D]$. To see it, set $x = 1$, and R cannot profitably deviate if, $k < \lambda[e_1 + \theta - c_R]$. Thus, R's incentive to deviate does not impose an additional parameter restriction.

Summing up, we've solved for a strategy profile for R and D that we can support as part of a PBE given $x \geq x^*$. While this is not the only strategy profile we can support, it is the one that can guarantee no agency loss at the lowest level of x. Thus, it is important to focus on it because (a) it allows Congress to avoid agency lost at the lowest $g(x)$, and (b) x only enters into Congress' payoff through $g(x)$. Thus, Congress strictly prefers $x = x^*$ over $x > x^*$.

Remark Suppose that in an equilibrium if Congress sets $x \geq x^*$, that Congress induces R and D to play the strategies described in Lemma A.6. Then Congress's expected utility at the moment Congress sets $x \geq x^*$ is:

$$EU^O(x \geq x^*) = pr[k < \lambda(e_1 + \theta - c_D)](\lambda(e_1 + \theta - c_O) - \mathbb{E}[k|k < \lambda(e_1 + \theta - c_D)]) - g(x) \quad (11)$$

A.9.4 Equilibrium

Define $\tilde{x}$ as the largest[36] x that maximizes:

[36]It may not be unique, but that is not the point. We pick the largest x because our goal is to show that D will pick one that is less than x^*.

$$\begin{cases} EU^O(x \geq x^*) & if\ x \geq x^* \\ EU^O(x < x^*) & if\ x < x^* \end{cases}$$

We now conjecture a set of strategies. O plays $\tilde{x}$. If $x \leq x^*$ R and D play the strategies written in proposition 3.3. If $x > x^*$, then R and D play the strategies described in Lemma A.6.

Proposition A.7 *Under our parameter restrictions, the conjectured strategies form a PBE.*

In section A.9.3 we showed that we could support R and D's strategy given an observed $x \geq x^*$. In section A.9.2 we argued via reference to proposition 3.3 that we could support R and D's strategy given an observed $x < x^*$. In the respective sections we defined O's expected utility from setting x given that it would induce the respective strategy. What is not proven is that O cannot profitably deviate from playing $\tilde{x}$, given the strategies for R and D it will induce. But by construction, $\tilde{x}$ is the x that (weakly) maximizes O's utility. Trivially O cannot deviate from it.

A.9.5 When will O set x to induce don't-ask-don't tell

To be clear, this result does not specify what $\tilde{x}$ is. It is possible that O would always set $\tilde{x} \geq x^*$. Our central claim is that conditions exist where we cannot support $\tilde{x} \geq x^*$. Thus, our final task is to verify that conditions exist where O will set $x < x^*$ and thus induce R and D to play the don't ask don't tell behavior.

We do so in two steps. First, we establish the best O can do if O sets x so large as to prevent don't-ask-don't-tell. Second, we establish conditions where O has at least 1 profitable deviation from O's best $x \geq x^*$.

Remark $EU^O(x \geq x^*)$ is maximized at $x = x^*$ for $x \geq x^*$. This yields:

$$EU^O(x = x^*) = pr[k < \lambda(e_1 + \theta - c_D)](\lambda(e_1 + \theta - c_O) - \mathbb{E}[k|k < \lambda(e_1 + \theta - c_D)]) - g(x^*) \quad (12)$$

Note that Congress's total expected utility for setting $x \geq x^*$ is weakly decreasing in x, because Congress must pay $g(x)$, But R and D's strategy are invariant to x, as are other features of Congress' utility. It follows that if Congress sets x to prevent don't ask don't tell, Congress sets $x = x^*$.

We now show conditions exist where Congress can profitably deviate from $x = x^* \to x = 0$. To be clear, this does not mean that $\tilde{x} = 0$. But it does guarantee that $\tilde{x} < x^*$, which is the point of our analysis. We focus on $x = 0$ because it simplifies the boundaries $\bar{k}$, allowing for a clear comparison. In particular, Congress's expected utility from $x = 0$ is:

$$EU^O(x = 0) = pr[k < \lambda(e_1 + \theta)]](\lambda(e_1 + \theta - c_O) - \mathbb{E}[k|k < \lambda(e_1 + \theta)]$$

For emphasis, we re-write it as:

$$EU^O(x=0) = EU^O(x=x^*) + g(x^*) \\ + pr[k \in [\underline{k}, \lambda(e_1+\theta)]] \times (\lambda(e_1+\theta-c_O) - \mathbb{E}[k|k \in [\underline{k}, \lambda(e_1+\theta)]]) \quad (13)$$

Remark In equilibrium, we cannot support $\tilde{x} \geq x^*$, on path if $EU^O(x=0) > EU^O(x=x^*)$:

$$g(x^*) > pr[k \in [\underline{k}, \lambda(e_1+\theta)]] \times (\mathbb{E}[k|k \in [\underline{k}, \lambda(e_1+\theta)]] - \lambda(e_1+\theta-c_O))$$

We note two facts about this inequality. First, if $g(x^*)$ is large, Congress will strictly prefer complete internal secrecy that induces don't ask don't tell to setting $x = x^*$. Thus, it instantly follows that the concern over external secrecy alone can drive Congress to set $x < x^*$.

But also notice that if we can ignore the direct costs by setting $g(x^*) = 0$ the inequality can still hold if:

$$\lambda(e_1+\theta-c_O) > \mathbb{E}[k|k \in [\underline{k}, \lambda(e_1+\theta)]]$$

The LHS of this inequality captures that lowering x from x^* to 0 means that research will happen leading to more innovation, and this raises the chance of welfare enhancing innovations. The RHS of this inequality captures that lowering x means that the additional research comes at a higher level of political costs.

B Monitoring the Soviets and the origins of U2

The main paper examined two cases of innovation: the search for mind control and the origins of the reconaissance satellite. This section examines a third case, the origins of the U-2 spy plane. As will be described in detail, this case provides additional inferential leverage that further validates the theory.

One of the United States' most pressing priorities in the early years of the Cold War was gaining better understanding of the Soviet Union's capabilities.[37] Without it, there was a heightened risk of insecurity, the possibility of arms racing, and even inadvertent war. But an aggressive and capable air defense made the prospect of overflights below a certain altitude a risky endeavor. Thus, the search for a high-flying reconnaissance aircraft was on.

The initial effort was spearheaded by the Air Force and various affiliated organizations. One of the most notable efforts was spearheaded by the Wright Air Development Command led by Major John Seaberg. In March 1953, Seaberg settled on desired specifications for the aircraft. He wanted it to "have an optimum subsonic cruise speed at altitudes of 70,000 feet or higher over the target, carry a payload of 100 to 700 pounds of reconnaissance equipment, and have a crew of one" (Pedlow and Welzenbach, 1992, 8). Seaberg solicited proposals from a number of smaller airframe manufacturing companies. He was seemingly interested in any solution that met his specifications and believed

[37] https://nsarchive2.gwu.edu/NSAEBB/NSAEBB74/U2-02.pdf.

smaller companies would take the project more seriously and move more quickly (Pedlow and Welzenbach, 1992, 8). He heard four bids:

- Fairchild Engine and Airplane Corporation proposed a single-engine aircraft, the M-195, which promised to reach a maximum altitude of 67,200 feet.
- Bell Aircraft Corporation proposed a twin-engine plane, the Model 67, or later the X-16, which promised to reach 69,500 feet.
- Glenn L. Martin Company proposed "a big-wing version of the B-57 called the Model 294, which was expected to cruise at 64,000 feet."
- Lockheed Aircraft Corporation proposed a modified, single engine aircraft that approximated sailplane, the CL-282, which promised to reach just north of 70,000 feet (Pedlow and Welzenbach, 1992, 9).

In a moment we will support our theory by examining who funds what and why. Before that, we emphasize the unique features of this case that help us validate our core counterfactual claim.

B.1 Counter-factual reasoning at this unique period in history

Our theory is built on a counter-factual claim: secret institutions pursue research that more open institutions would reject. This is difficult to validate in the modern institutional context for three reasons. First, the military and intelligence organizations employ many scientists who devise ideas on their own. When a CIA scientist conceives of a novel idea and explores it, for example, we cannot know whether the military would have rejected it. Second, scientists and engineers select into the institutions they work for. As such, we cannot know if CIA scientists are similar to military scientists and vice versa. Finally, private companies that devise new ideas know they can pitch them to highly secret parts of the government like the CIA through classified contract mechanisms. If our theory is right, we may never observe them take ideas to the military.

A confluence of factors in this case provides a unique opportunity to test our theory. First, the companies that bid on reconnaissance aircraft all believed that the Air force was effectively the sole outlet for such pitches.[38] Interestingly, however, a relevant secret organization did exist. In July 1954, President Eisenhower tapped the President of MIT, James Killian, to head a group of scientific experts called the Technology Capabilities Panel (TCP) (Richelson, 2002, 11). Its existence was not widely known: "As with other secret panels formed by chief executives to deal with intelligence matters, Congressional input was missing from the TCP deliberations and few Congressmen knew it even existed, although many of its decisions had an immense impact on the nation's military and intelligence preparedness" (Laurie, 2001, 5).

Project Three, one of three entities comprising the TCP, was a small group broadly focused on intelligence capabilities. It was not specifically tasked with developing proposals for overhead reconnaissance aircraft. Thus, the small and secretive Project Three members were not soliciting bids for such aircraft, and nobody expected that they would. However, the extreme secrecy that

[38] Although the CIA had developed several branches to deal with scientific intelligence and research and development in the early- to mid-1950s, they did not have much experience at that time with technical collection systems. See Fischer (2001).

surrounded Project Three meant that they could develop research ideas in small teams that outsiders would not know about. Thus, unlike the Air Force, they exhibit the internal secrecy that our theory requires for secret innovation.

Based on this context, it is reasonable to assume that the Wright Air Development Command and any other relevant Air Force-related entity would hear all bids pertaining to overhead reconnaissance and had first right of refusal. Moreover, any project they did fund would at least be scrutinized by the broader Air Force leadership and possibly Congress. They would have also likely believed that anything they rejected would not be funded. However, as just noted, Project Three was quietly lurking in the background and ready to pick up rejected proposals if they so chose. This allows us to evaluate our counterfactual because we can observe: (1) what the open institution actually chose to accept and reject and; (2) given what the open institution rejected, what the secret institution chose to accept and reject.

B.2 Who funds what and why

The Air Force opted to pursue two proposals, the modified version of the B-57 from Martin which was viewed as a short-term solution and the Bell X-16 which promised better results in the medium-term. Bell was contracted to produce 28 such aircraft. At the same time, the Air Force rejected the Fairchild and Lockheed proposals. The Fairchild proposal was relegated to the dustbin of history. The Lockheed proposal was not. Lockheed took their proposal to various parts of the Air Force—including the Wright Air Development Command as well as Strategic Air Command and the Office of Development Planning—all of whom rejected it (Pedlow and Welzenbach, 1992, 11-12). Along the way, Project Three members learned of the Lockheed proposal and were immediately interested in it (Pedlow and Welzenbach, 1992, 31). As we will detail more in a moment, they undertook intense secretive research into CL-282's viability and verified that it would work. This project was later handed to the CIA as the U-2 project.

We predict that open organizations facilitate innovation when the benefits are clear (e_0 is positive), and there is not much disagreement about the likely effects (σ_0 is low); they will reject ideas that are radically new because they know little about them. Even though new ideas could have benefits, they could also cause damage. Open institutions are unlikely to take on projects like this even in the research phase (e_0 is near 0). Of these ideas, we predict that secretive institutions will pick them up as research projects if the potential outcomes vary widely (σ_0 is high). That is, there is a risk of catastrophic damage towards mission objects and enormous benefits that extend beyond what the other proposals could accomplish.[39]

This is precisely what we find. The Air force funded two safer projects that incrementally advanced the state of overflight. The modified B-57 is an obvious example. The goal was to "improv[e] the already exceptional high-altitude performance of the B-57 Canberra" (Pedlow and Welzenbach, 1992, 9). It "featured lengthened wings, accommodations for cameras and sensors, and uprated twin engines" (Merlin, 2015, 1). The Bell X-16 was slightly more advanced than the B-57. The modifications made to reduce weight and reach higher altitudes were far less radical than the CL-282 (Merlin, 2015, 4-5).

The U-2 was radical by design. Senior Lockheed designers prioritized "nonstandard" elements,

[39] That is, there is uncertainty about whether the innovation will move the U.S. closer to or further from its policy objectives.

including "the elimination of landing gear, the disregard for military specifications, and the use of very low load factors" (Pedlow and Welzenbach, 1992, 10). Several elements of what was eventually dubbed the CL-282, and would later become the U-2, "were adapted from gliders. Thus, the wings and tail were detachable. Instead of conventional landing gear," Kelly Johnson, the lead developer, "proposed using two skis and a reinforced belly rib for landing—a common sailplane technique—and a jettisonable wheeled dolly for takeoff." As a declassified history of the U-2 puts it, "Essentially, Kelly Johnson had designed a jet-propelled glider" (Pedlow and Welzenbach, 1992, 12).

Part of Seaberg and the Wright Air Development Command's rationale for rejecting the CL-282 proposal speaks to their uncertainty about whether it would work. Seaberg pointed to its use "of the unproven General Electric J73 engine. The engineers at Wright Field considered the Pratt and Whitney J57 to be the most powerful engine available." All three of the other proposals they received from small manufacturers relied on the latter. Moreover, Seaberg and colleagues viewed "[t]he absence of conventional landing gear" on the CL-282 as a "shortcoming." Because the other proposals, including the most promising—the Bell—had "normal landing gear," they were considered "more conventional aircraft" (Pedlow and Welzenbach, 1992, 12-15).

Other Air Force commands also registered dismay at the novel features of CL-282. General Curtis LeMay, the head of Strategic Air Command, apparently "stood up halfway through the briefing, took his cigar out of his mouth, and told briefers, that if he wanted high-altitude photographs, he would put cameras in his B-36 bombers and added that he was not interested in a plane that had no wheels or guns." He called the meeting "a waste of his time" (Pedlow and Welzenbach, 1992, 12).[40]

According to the declassified history of the U-2, another driving factor in the Air Force's rejection of the CL-282 had to do with their "preference for multi-engine aircraft." This was based on familiarity and their experience with multi-engine aircraft during World War II and likely explains why they also opted for the Bell and Martin designed but rejected the Fairchild bid, which relied on a single engine. Moreover, "aerial photography experts" at the time "emphasized focal length as the primary factor in reconnaissance photography and, therefore, preferred large aircraft capable of accommodating long focal-length cameras" (Pedlow and Welzenbach, 1992, 13)

As the foregoing makes clear, the CL-282's novel design meant that many in the Air Force were skeptical about its chances of success. In terms of the model's parameters, the balance of Air Force staff thought the overall impact of the project would cause no benefit (or harm) for surveilling the Soviet Union and ultimately ensuring peace. However, some raised concerns which implied that it could have catastrophic effects: "there was the feeling shared by many Air Force officers that two engines are always better than one because, if one fails, there is a spare to get the aircraft back to base... Furthermore, a high-altitude reconnaissance aircraft deep in enemy territory would have little chance of returning if one of the engines failed, forcing the aircraft to descend" (Pedlow and Welzenbach, 1992, 13). In other words, there was concern that a single-engine plane that was missing key parts could crash inside the Soviet Union and conceivably spark a conflict.

To be sure, not everyone in the Air Force shared the view that the Bell and Martin proposals were superior to the CL-282. Trevor Gardner, Special Assistant for Research and Development, and some other officials thought it had potential. They believed "it gave promise of flying higher than the other designs and because at maximum altitude its smaller radar cross-section might make it

[40] LeMay's reaction illustrates one way that military culture imposes costs on innovators. As we argued, this makes innovation difficult in open institutions.

invisible to existing Soviet radars" (Pedlow and Welzenbach, 1992, 15). Thus, if it worked, its value would be larger than the other projects.

Taken together, these divergent views support the notion that there was deep uncertainty about what CL-282 would accomplish. While some believed it was unlikely to work and therefore have no effect, others thought it could have either very negative or very positive (i.e. more positive than the other designs) effects. If the Air Force had been the only organization that could have considered the overflight proposals, one of the most important innovations of the twentieth century may never have seen the light of day (Pocock, 2000, 14).

Project Three members were themselves sensitive to the risks associated overflight over the Soviet Union.[41] But despite these risks, they pursued the project because of the enormous potential upside if the project was successful. "By the end of October [1954], the Project Three meetings had covered every aspect of the Lockheed design. The CL-282 was to be more than an airplane with a camera, it was to be an integrated intelligence-collection system that the Project Three members were confident could find and photograph the Soviet Union's Bison bomber fleet and, thus, resolve the growing 'bomber gap' controversy." They were also taken with the prospect that the proposal could be "the platform for a whole new generation of aerial cameras" (Pedlow and Welzenbach, 1992, 31).

Their approach to research supports our theory in two additional ways. First, they operated in secret. Land and his team "began developing it into a complete reconnaissance system," meeting in small-group settings with usually less than 10 people present. Second, they did not instantly recommend production of U-2 planes. Rather, they exploited secrecy to determine if the project was viable. Once they realized it was, they revealed what they had been doing to the CIA Director and to President Eisenhower who was extremely receptive. He "approv[ed] the development of the system, but ... stipulat[ed] that it should be handled in an unconventional way so that it would not become entangled in the bureaucracy of the Defense Department or troubled by rivalries among the services" (Pedlow and Welzenbach, 1992, 33).[42]

Interestingly, the project also helped the TCP realize that secret organizations like the CIA were well-suited to the task of overseeing radical innovations of this kind. As the TCP argued to CIA Director Allen Dulles in a memo, "this seems to us the kind of action and technique that is right for the contemporary version of the CIA; a modern and scientific way for an Agency that is always supposed to be looking, to do its looking. Quite strongly, we feel that you must always assert your first right to pioneer in scientific techniques for collecting intelligence... This present opportunity for aerial photography seems to us a fine place to start" (Land, 1954*b*).

C National Security and Innovation Literature

Since our theoretical framework is closest to principal-agent theories of organizational innovation, we focus our review on that literature. We also review works in international relations and bureaucratic politics that help us justify changes in our assumptions. However, our paper has broad substantive interest for scholars of innovation and security broadly defined. Here we review four different

[41] See Land (1954*a*).

[42] Interestingly, the Air Force eventually comes around to accepting the proposal but does not actually abandon their X-16 program until the U-2 was operational.

strands of this literature, explain how we connect and contribute to them:

1. Bureaucracy and barriers to and opportunities for military innovation;
2. Adaptation and military innovation;
3. Conflict processes and innovation, which can examine autocratic repression or terrorism and innovation;
4. The strategic implications of new technology.

Many of the concepts we describe intersect with these literatures. But we frequently arrive at surprising conclusions for all of them. In what follows, we explain how our theory intersects with these important literatures and clarify differences.

C.1 Barriers and opportunities for military innovation

A large literature in security and strategic studies examines military innovation. Many of these analyses begin with the premise that, despite the importance of innovation to national security, military innovation is rarer than we might expect it to be. Why? The answer, in brief, is that innovators face costs of different kinds. One common impediment is that militaries are "hierarchical, inflexible, and rigid" (Jungdahl and Macdonald, 2015, 467). As Grissom (2006, 919) argues in his review of this literature, most scholars argue that "military organizations are intrinsically inflexible, prone to stagnation, and fearful of change." What this means in practice is that individuals are often reluctant to suggest new ideas for professional or cultural reasons, and new ideas that do get proposed can often get shut down.

Despite these barriers, militaries sometimes innovate. Thus, another key task of this literature is to answer the following question: what explains how militaries can overcome bureaucratic inertia or military culture to innovate? Some argue that military organizations may innovate when they face external pressures from the outside, usually from civilians (Posen, 1984). Another is when senior members of the military re-conceptualize their tasks and create career paths for new officers that incentivize the embrace of this new way of thinking (Rosen, 1988). A third set of explanations focuses on cultural differences (Adamsky, 2010; Farrell and Terriff, 2002) According to one study, a "receptive culture" can facilitate new thinking and vice versa.[43] A fourth argues that innovation requires special incubators where individuals can collaborate, try out ideas, and push the envelope Jensen (2016). There are others (Grissom, 2006).

While each of these pathways are distinct in important ways, they all share a common strategic logic. First, individuals inside the military face barriers (i.e. costs) to innovation. Therefore, they either do not voice their ideas, or are unable to push their ideas through the military bureaucracy. This explains why innovation does not happen often. Second, opportunities for innovation arise when military leaders, or outsiders with power create incentives (i.e. lower the costs associated with pursuing innovation). Things like new pathways to promotion, visionary civilians that intervene to support and defend new ways of doing business, and incubators where individuals can test

[43] Price (2014). Lee (2019) has shown, for example, that the Air Force's cultural preference for manned systems led it to reject innovations in drone technology for longer than would otherwise be the case if one were using a strictly rationale model.

ideas outside the formal process are a way for would-be innovators to safely conceive of ideas, develop them, and potentially implement them without incurring significant costs. Without these cost-lowering mechanisms, the argument goes, innovation does not happen.

Our theory accounts for these conditions in the costs and benefits parameters. The logic of our model under a specific set of parameters is consistent with the logic of these arguments. We find that researchers will not *openly* pursue innovation even when the policy implications are important (the expectation of pi is positive) if the organization imposes large personal costs on the agents.

The critical difference between our theory and this literature is what happens when the costs and benefits are high. Scholars of military innovation typically argue that if the costs of pursuing research are high then the innovators simply do not pursue their ideas. As noted, their logics for military innovation largely follow a similar process: some kind of organizational change transpires that lowers the costs associated with agents openly pursuing innovation; the researcher realizes that the organization is accommodating of new ideas; the researcher then raises their ideas with their manager so that they can openly pursue them. In our theory, national security researchers sometimes face another option: secret innovation. Rather than taking their idea to their manager, or sharing it broadly with others in their organization, a small team of researchers can pursue an idea in secret. This gives the researcher autonomy to pursue their idea and demonstrate its plausibility. It also allows different agents to distribute the high institutional costs associated with pursuing new ideas.

In this way, our theory illuminates that existing studies emphasize open, national security innovation in the way that we define openness.[44] As written in the manuscript, open research refers to a setting where individuals broadly share their ideas with their managers, people with budgetary oversight, and many others across their organization and possibly outside their organization.[45] What is more the costs that these scholars describe usually stem from openness. Consider that bureaucratic inertia, or cultural barriers only prevent pilot testing if ideas are shared openly. If a small team of researchers does not ask permission, they do not face bureaucratic inertia.

There are several other ways in which our theory differs from, but complements, broader literature on military innovation which includes both doctrinal innovations as well as technological and tactical innovations (Beard, 1976; Jungdahl and Macdonald, 2015; Sapolsky, 1972) First, most of these accounts emphasize innovation that occurs through a top-down process. Our focus entails a heavy bottom-up component (Griffin, 2017, 214).[46] Second, much of this scholarship on military innovation has a bias towards *successful* innovations.[47] By focusing on the process or pursuit of innovation, our study allows for the prospect that many of these ideas, particularly those pursued

[44] Scholars such as Kurth Cronin (2020, 23-28) discuss "closed innovation," defined as "state organizations creat[ing] and control[ing] high-end military technologies" such as nuclear weapons. Even in this case, though, while innovation may be hidden from the *outside* world it is still open internally within the government.

[45] Although they do not usually describe it this way, the existing security studies literature usually focuses on open innovation under this definition. Perhaps the clearest example of this is innovations in doctrine, a common focus of this literature. When doctrinal innovation happens, it is usually carried out in broad view of many parts of the military. It requires many services and branches to work together. Even during periods of conceptualization, new doctrine requires combat experts to interface with logistics, strategic intelligence, manpower and budget experts, defense contractors, and more. Moreover, since new doctrine requires new field manuals, soldiers tend to find out important details of doctrine as it is being developed.

[46] For exceptions, see Jungdahl and Macdonald (2015); Kollars (2014).

[47] This is evidenced by the way many scholars define innovation, which often requires things like improvements in military effectiveness. See Grissom (2006, 907). As Posen (1984, 29) notes, however, "Neither innovation nor stagnation ... should be valued a priori.

in secret organizations, will fail.

C.2 Adaptation and military innovation

A second literature examines diffusion and adaptation. This is similar because it examines military innovation. However, they focus on how existing military technologies diffuse cross-nationally. Horowitz (2010, 3), for example, develops the "adoption-capacity theory" to explain "why some military innovations spread and influence international politics while others do not, or do so in very different ways." In a somewhat similar vein, Gilli and Gilli (2019, 141) examine the logic of imitation, asking whether America's rivals can "easily imitate its most advanced weapon systems and thus erode its military-technological superiority."

The aspect of these studies that is most similar to ours examines different ways that states adopt the same technology. This could be thought of as tactical innovations. However, these tactical innovations are typically described as open, and the primary barriers is in adopting an existing technology and not in finding new ways to use it.

C.3 Innovation among autocrats and terrorist groups

Our framework also differs from a newer literature on innovation among terrorist organizations and autocratic regimes. Regarding terrorist groups, innovation is often driven by the need to evade a target's defenses, amplify lethality, and shape public opinion (Horowitz, Perkoski and Potter, 2018). The precise characteristics of terrorist organizations, their leaders, and their broader environment, however, shape whether they are successful.[48] One of the key differences between these studies and our own is that terrorist organizations as a whole are insensitive to the costs of innovation whereas the individuals in our model are political actors and researchers with an entirely different incentive structure.[49]

Finally, there is an emerging literature that examines innovation and autocratic regimes. A key focus of these works is how dictators can exploit technological innovations to their advantage. This includes the use of the Internet and other technologies for the purposes of repression and surveillance (Dragu and Lupu, 2021; Gohdes, 2020). In these studies, autocratic leaders are exploiting existing technologies that may have been developed with an entirely separate purpose in mind for their own ends, including regime survival and population control. Like terrorist organizations, they are also insensitive to costs. As noted, our focus is on the sources of innovation in a situation where there are political actors who can distribute costs to subordinates.

C.4 Strategic implications of emerging technology

A growing litreature emphasizes the strategic implications of emerging technology (see Sechser et al., 2019, for review). We partly use this litearture to justify our claim that the benefits of innovation (i.e. whether innovation moves you closer or further from your policy goals) is uncertain. This

[48] See Moghadam (2013); Perkoski (2019).

[49] To be sure, terrorists may be sensitive to how the public will *perceive* an innovation such as suicide bombing but are themselves by and large insensitive or at least willing to incur enormous costs given the nature of asymmetric conflict.

literature is more about what states do with innovations once they have them. It is less about why states decide to pursue them in the first place (Garfinkel and Dafoe, 2019; Horowitz, 2016; Zhang et al., 2021).

D Principal-agent literature

Our substantive focus is foreign policy and international relations. However, as we discuss in the manuscript, the structure of our theory is closest to principal-agent theories of organizational innovation in the private sector (Lai et al., 2009). These theories emphasize aspects of PA problems not commonly studied by international relations scholars. In what follows, we explain how our theory fits within the PA framework. We then clarify important differences with three applications of PA theory in IR.

D.1 What makes our theory a principal-agent theory?

PA theory is very broad (Eisenhardt, 1989). There are many types of principal-agent problems that scholars study including moral hazard, agency loss, adverse selection, credible communication, and unjust reprisals (Stiglitz, 1989; Hart and Holmström, 1987). While each problem is different, they are united by a few common elements. In this section, we describe the elements of a PA theory and how our theory includes these elements.

A basic principal-agent dynamic (or contract theory) involves at least one agent and at least one unified principal that have asymmetric preferences and in which the agent is given a choice to impact the principal's welfare (Miller, 2005). Our basic institution models these elements. We study a researcher and manager who vary in their cost functions. As a result of these cost functions, situations arise where the researcher wants to pursue research and development but the manager does not. We make one assumption that is common in models of innovation: the effects of pursuing a policy follow from imperfect information and are not known to either player. This assumption is not common in PA models of policymaking (e.g. Downs and Rocke, 1994). The reason is that policymakers (i.e. the agent) knows whether their choice will benefit the principal with a large degree of confidence (i.e the public); at least *ex post*.

Beyond this difference, we make a novel assumption in the basic model that departs from PA models of innovation: the researchers can exploit secrecy to distribute costs. This creates a dynamic in which the researcher can incur costs to pursue outcomes that the manager would veto. We study the impact of this additional assumption under complete information because it generates a novel tension not typically appreciated in PA models.

Principal-agent theories introduce problems through asymmetric information, and a principal's initiative (Miller, 2005). The specific type of principal-agent problem varies depending on how scholars introduce private information (Hart and Holmström, 1987). We model two variants of a principal-agent problem in extensions 3.3.1 and 3.3.2. The first represents a monitoring problem, the second represents a credible advice problem. Past scholars examine how variation in the costs of monitoring, agent selection, or punishments can elicit agency compliance and the credible revelation of information. However, we find that secrecy paradoxically alleviates many of the common problems of asymmetric preferences and information. It also creates new incentives for the manager

to extract value from the researcher's compliance.

D.2 How is this different from PA models in international relations and foreign policy studies?

Here we describe three literatures that examine principal-agent problems in international relations: hierarchy, security force assistance, and gambling for resurrection.

We start with a joint-discussion of hierarchy (Hawkins et al., 2006; Nielson and Tierney, 2003; a. Lake, 2001) and security force assistance (Biddle et al., 2018; Ladwig, 2016). Of course, these empirical domains are very different from each other. Further, each domain includes many different studies that tackle different aspects of the PA problem. However, they are all united by the fact that they assume the principal and agent come from different states and therefore have dramatically different preferences. Scholars of security force assistance assume that the principal is either US military advisers or the entire US military and the agent is the military of another state (e.g. the Afghan army).

We do not focus on a situation like this. Consistent with organizational models of principal-agent theory and innovation, we examine individual employees (or small groups of individuals) who work at a single organization (or a handful of closely connected agencies that share a common mission within the executive branch of a single country; like the CIA and NRO). To match this domain, we assume that the researcher and manager both share an interest in advancing the organization's overall goals (both researcher and manager's utility is increasing in π). However, their preferences over research and development still vary because the personal and professional incentives of managers and researchers vary (c, k can vary).

Our assumptions are appropriate for the setting we study. The goals that national security agencies pursue are things like defeating the Soviet Union in the Cold War, or winning the Second World War. In general, we believe that managers and researchers employed in the national security community benefit to the extent that they succeed in these goals and lose to the extent they fail in them. This is partly due to the extensive security clearance process and constant monitoring that national security employees are subject to. It also relates to professional incentives once in these communities. Finally, evidence suggests that public-sector employees, and especially national security employees, tend to have a strong public service motivation. However, individual agents may disagree about the best way to achieve these goals, face incentives to buck-pass, or have parochial incentives that cause them to weight the costs and benefits differently.

Studies of gambling for resurrection are closer to us because they examine a leader and the public of the same country. Most notably, Downs and Rocke (1994) theorize about the president as the agent who makes the choice to fight a war (or not). The president holds asymmetric information over whether war serves the public interest. They model the public as the principal who can re-elect the president. This model is closer to ours than the hierarchy and security force assistance literatures in that the public and the president both share a preference for avoiding bad foreign policy outcomes.

But there are several differences. First, the president has a unique incentive for re-election that can conflict with the public's. As discussed, these preferences are not appropriate in our theory (although our theory is robust if we model preference variation like this). Second, the president

has private information about the quality of the choice to fight, and his own quality. This is not appropriate in our model for two reasons. The first reason is, unlike the American public, the manager has a security clearance and access to a wide cadre of classified researchers who can review the existing data. The second reason is that the researcher is very uncertain before they engage in pilot research precisely because they have not worked on a problem like this. Third, the public directly punishes the president through an electoral mechanism. This is not appropriate in our theory for two reasons. One reason is that the manager is complicit through don't-ask-don't tell, and therefore does not do the punishing. Another is that punishment does not take the form of replacing a researcher with a different one (as in the electoral context).

References

a. Lake, David. 2001. "Beyond Anarchy: The Importance of Security Institutions." *International Security* 26:129–160.

Adamsky, Dima. 2010. *The Culture of Military Innovation: The Impact of Cultural Factors on the Revolution in Military Affairs in Russia, the US, and Israel*. Stanford: Stanford University Press.

Beard, Edumnd. 1976. *Developing the ICBM: A Study in Bureaucratic Politics*. New York: Columbia University Press.

Biddle, Stephen, Julia Macdonald and Ryan Baker. 2018. "Small footprint, small payoff: The military effectiveness of security force assistance." *Journal of Strategic Studies* 41:89–142.

Downs, George W. and David M. Rocke. 1994. "Conflict, Agency, and Gambling for Resurrection: The Principal-Agent Problem Goes to War." *American Journal of Political Science* 38:362.

Dragu, Tiberiu and Yonatan Lupu. 2021. "Digital Authoritarianism and the Future of Human Rights." *International Organization* 75(4):991–1017.

Eisenhardt, Kathleen M. 1989. "Agency Theory: An Assessment and Review." *The Academy of Management Review* 14:57.

Farrell, Theo G. and Terry Terriff. 2002. *The sources of military change: Culture, politics, technology*. Boulder, CO: Lynne Rienner.

Garfinkel, Ben and Allan Dafoe. 2019. "How does the offense-defense balance scale?" *Journal of Strategic Studies* 42:736–763.

Gilli, Andrea and Mauro Gilli. 2019. "Why China Has Not Caught Up Yet: Military-Technological Superiority and the Limits of Imitation, Reverse Engineering, and Cyber Espionage." *International Security* 43(3):141–189.

Gohdes, Anita R. 2020. "Repression technology: Internet accessibility and state violence." *American Journal of Political Science* 64(3):488–503.

Griffin, Stuart. 2017. "Military Innovation Studies: Multidisciplinary or Lacking Discipline?" *Journal of Strategic Studies* 40(1-2):196–224.

Grissom, Adam. 2006. "The future of military innovation studies." *Journal of strategic studies* 29(5):905–934.

Hart, Oliver and Bengt Holmström. 1987. *The theory of contracts*. Cambridge University Press pp. 71–156.

Hawkins, D G, D A Lake, D L Nielson and M J Tierney. 2006. *Delegation and Agency in International Organizations*. Cambridge University Press.

Horowitz, Michael. 2010. *The diffusion of military power : causes and consequences for international politics*. Princeton University Press.

Horowitz, Michael C. 2016. "Public Opinion and the Politics of the Killer Robots Debate." *Research & Politics* 3(1):1–8.

Horowitz, Michael C., Evan Perkoski and Philip B.K. Potter. 2018. "Tactical Diversity in Militant Violence." *International Organization* 72(1):1–35.

Jensen, Benjamin M. 2016. *Forging the Sword: Doctrinal Change in the U.S. Army*. Stanford, CA: Stanford University Press.

Jungdahl, Adam M and Julia M Macdonald. 2015. "Innovation inhibitors in war: Overcoming obstacles in the pursuit of military effectiveness." *Journal of Strategic Studies* 38(4):467–499.

Kollars, Nina. 2014. "Military innovation's dialectic: Gun trucks and rapid acquisition." *Security Studies* 23(4):787–813.

Kurth Cronin, Audrey. 2020. *Power to the People: How Open Technological Innovation is Arming Tomorrow's Terrorists*. New York: Oxford University Press.

Ladwig, Walter C. 2016. "Influencing Clients in Counterinsurgency: U.S. Involvement in El Salvador's Civil War, 1979–92." *International Security* 41:99–146.

Lai, Edwin L.-C., Raymond Riezman and Ping Wang. 2009. "Outsourcing of innovation." *Economic Theory* 38:485–515.

Lee, Caitlin. 2019. "The Role of Culture in Military Innovation Studies: Lessons Learned from the US Air Force's Adoptino of the Predator Drone, 1993-1997." *Journal of Strategies Studies* pp. 1–35.

Miller, Gary J. 2005. "THE POLITICAL EVOLUTION OF PRINCIPAL-AGENT MODELS." *Annual Review of Political Science* 8:203–225.

Moghadam, Assaf. 2013. "How al Qaeda innovates." *Security Studies* 22(3):466–497.

Nielson, Daniel L. and Michael J. Tierney. 2003. "Delegation to International Organizations: Agency Theory and World Bank Environmental Reform." *International Organization* 57:241–276.

Perkoski, Evan. 2019. *Terrorist technological innovation*. Oxford: Oxford University Press.

Posen, Barry. 1984. *The sources of military doctrine: France, Britain, and Germany between the world wars*. Ithaca: Cornell University Press.

Price, John F. 2014. "US Military Innovation: Fostering Creativity in a Culture of Compliance." *Air & Space Power Journal* 43(Sep.-Oct.):128–134.

Rosen, Stephen Peter. 1988. "New ways of war: understanding military innovation." *International security* 13(1):134–168.

Sapolsky, Harvey M. 1972. *Polaris System Development: Bureaucratic and Programmatic Success in Government*. Cambridge, MA: Harvard University Press.

Sechser, Todd S., Neil Narang and Caitlin Talmadge. 2019. "Emerging technologies and strategic stability in peacetime, crisis, and war." *Journal of Strategic Studies* 42:727–735.

Stiglitz, Joseph E. 1989. *Principal and Agent*. Palgrave Macmillan UK pp. 241–253.

Zhang, Baobao, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz and Allan Dafoe. 2021. "Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers." *Journal of Artificial Intelligence Research* 71:591–666–591–666.