Inaugural-Dissertation

zur Erlangung der Doktorwürde der Wirtschaftswissenschaften

des Fachbereichs Wirtschaftswissenschaften

der Philipps-Universität Marburg

Incentives for Researchers

vorgelegt von

Matthias Verbeck, M.A.

aus Hachenburg

Erstgutachterin:	Prof. Dr. Elisabeth Schulte
Zweitgutachterin:	Prof. Dr. Evelyn Korn
Prüfungsvorsitzender:	Prof. Dr. Tim Friehe
Einreichungstermin:	6.12.2021
Prüfungstermin:	23.2.2022
Erscheinungsort:	Marburg
Hochschulkennziffer:	1180

Danksagung

An dieser Stelle möchte ich mich von Herzen bei denjenigen Personen bedanken, die mich während der Zeit meiner Promotion wissenschaftlich oder persönlich begleitet und somit zum Gelingen des Projekts maßgeblich beigetragen haben.
Mein größter Dank gilt meiner Promotionsbetreuerin Prof. Dr. Elisabeth Schulte, die mich mit viel Einsatz, Geduld und konstruktiven Hinweisen bei der Anfertigung der vorliegenden Arbeit unterstützt hat. Zudem danke ich ihr für die vielen Einsichten und Erkenntnisse, die ich durch Zusammenarbeit und Gespräche während meiner Zeit als wissenschaftlicher Mitarbeiter in der Abteilung Institutionenökonomie gewinnen durfte. Aus dem Professorium der Universität Marburg bedanke ich mich zudem herzlich bei Prof. Dr. Evelyn Korn, insbesondere für den von ihr geleiteten Ph.D-Kurs in Spieltheorie, sowie bei Prof. Dr. Bernd Hayo für viele konstruktive Anregungen in internen Forschungsseminaren bedanken.
Aus dem Kreis der (ehemaligen) Doktorandinnen und Doktoranden und des wissenschaftlichen Personals am Fachbereich 02 gilt mein Dank besonders Dr. Enrico Böhme und Dr. Christoph Rößler für viele anregende und unterhaltsame Gespräche und die notwendige Zerstreuung neben der Arbeit. Auch für den freundschaftlich-kollegialen Austausch und die gute Zusammenarbeit mit Philipp Bösherz, Isabel Cutrim, Dr. Severin Frank, Dr. Helge Müller, Dr. Volker Robeck, Prof. Dr. Dr. Hannes Rusch, Dr. Matthias Uhl, Dr. Johannes Zahner und Dr. Johannes Ziesecke möchte ich herzlich danke sagen.
Ebenso bedanke ich mich bei allen ehemaligen Kolleginnen und Kollegen in der Abteilung Institutionenökonomie, insbesondere bei Kai Brenneke, Max Burger, Bärbel Dönges, Rebecca Dietrich, Jonathan Gehlen, Adrian Pourviseh, Malin Olivia Soeder und Verena Schröder.
Abschließend möchte ich mich herzlich bei meiner Familie und bei meinen Freunden, insbesondere bei Philipp Klevers, Sven Otto, Fabiane Engelmann und Jana Pohl bedanken, die mich während meiner Promotionszeit stets mit Rat und Tat unterstützt haben.

Ihnen und Euch allen vielen Dank!

Dokumentation der Ko-Autorenschaft

Kapitel 1 “Contracting with Researchers” wurde von Prof. Dr. Elisabeth Schulte und mir verfasst. Die Konzeption des Modells stammt von mir. Die Analyse des Modells wurde unterstützend von Prof. Dr. Elisabeth Schulte und mir durchgeführt.

Kapitel 2 “Strategic Delay in R&D Projects - An Agency Perspective” wurde ebenso in gemeinsamer Autorenschaft verfasst. Die Forschungsidee stammt von mir, wobei die Forschungsfrage im Laufe der Arbeiten neu ausgerichtet wurde. Die finale Version (inklusive der Entwicklung und Analyse des Modells) wurde zu gleichen Teilen von Prof. Dr. Elisabeth Schulte und mir erarbeitet.

Kapitel 3 “The Inspection Game in Science” wurde in Alleinautorenschaft von mir verfasst. Hierbei wurde ich durch regelmäßiges Feedback von Prof. Dr. Elisabeth Schulte unterstützt.

Inhaltliche Zusammenführung

"*Der Wissenschaftler findet seine Belohnung in dem, was Henri Poincaré die Freude am Verstehen nennt, nicht in den Anwendungsmöglichkeiten seiner Entdeckung.*" (Albert Einstein, zitiert nach Planck (1932), S. 211, Übersetzung durch den Autor). Dem obigen Zitat von Albert Einstein aus dem Jahr 1932 würden sicherlich auch heute noch viele Wissenschaftlerinnen und Wissenschaftler zustimmen. Die Freude an der Entdeckung neuer Erkenntnisse und am Verstehen bisher nicht bekannter Zusammenhänge dürfte eine der Hauptantriebskräfte für viele der im Wissenschaftsbetrieb tätigen Forscherinnen und Forscher sein.

Gleichwohl ist das von Einstein gezeichnete Bild eines ausschließlich vom Forscherdrang und Erkenntnisgewinn getriebenen Wissenschaftlers - zumindest aus der Sicht der Wirtschaftswissenschaften - ein idealisiertes. In der akademischen Realität sind Forschungsgelder und Professuren knapp, der Platz in hochrangigen akademischen Zeitschriften ist begrenzt und die Anerkennung der Kolleginnen und Kollegen beschränkt sich zuvorderst auf jene wenigen unter ihnen, welche mit eindrucksvollen Resultaten aufwarten können. Aus Sicht der Institutionenökonomik, welche die Wichtigkeit von Anreizen betont, steht daher zu vermuten, dass die Freude an der Forschung an sich nicht das einzige Handlungsmotiv von Wissenschaftlerinnen und Wissenschaftlern darstellen kann, und beispielsweise Karrieremotive ebenso eine Rolle spielen. Der akademische Betrieb bildet hier den institutionellen Rahmen, innerhalb dessen Forschende ihr Verhalten so anpassen, dass ihr individueller Nutzen maximiert wird. Dabei wird - wenig idealistisch - angenommen, dass auch vor opportunistischem Verhalten nicht zurückgeschreckt wird. Dadurch können beispielsweise Wissenschaftsbetrug oder andere fragwürdige wissenschaftliche Praktiken wie "p-Hacking"erklärt werden, welche in Einsteins Bild keinen Platz finden.

Die Möglichkeiten zu opportunistischem Verhalten sind in der Forschung besonders ausgeprägt, da diese strukturell von Informationsasymmetrien gekennzeichnet ist. Sowohl die erzielten Forschungsergebnisse als auch die verwendeten Methoden und der eingesetzte Forschungsaufwand sind im Regelfall die private Information des jeweiligen Forschenden und nur bedingt durch Dritte verifizierbar. Es verwundert also nicht, wenn beispielsweise Jean Tirole schreibt, dass die Beziehung zwischen Forschenden und ihren Finanzierungsquellen "von Moral Hazard durchzogen" sei (Tirole (2006), Übersetzung durch

den Autor). Zugleich trägt die Arbeit von Forscherinnen und Forschern maßgeblich zu technologischem Fortschritt und dem Entstehen von Produktinnovationen bei und ist somit für das Gedeihen und das Wachstum von Volkswirtschaften von enormer Wichtigkeit. Etwaige Friktionen und Fehlanreize innerhalb der Wissenschaft haben somit direkte Auswirkungen auf die Gesamtwohlfahrt.
Die vorliegende Dissertation hat daher zum Ziel, die Anreize und Institutionen in der (akademischen) Forschung besser zu verstehen. Sie besteht aus drei eigenständigen Aufsätzen, die sich - jeweils aus institutionenökonomischer Perspektive mithilfe modelltheoretischer Analysen - mit je einem spezifischen Problemfeld aus dem Kontext akademischer bzw. industrieller Forschung beschäftigen.

Der erste Aufsatz thematisiert die individuelle Auswahl von Forschungsansätzen bzw. Forschungstechnologien durch die Forschenden im Kontext *delegierter* Forschung. Es wird gezeigt, dass es hierbei zu einer unzureichenden Diversität bei der Auswahl der verwendeten Forschungstechnologien kommen kann. Analysiert wird ein Modell mit zwei risikoaversen Agenten (den Forschenden), die im Auftrag eines risikoneutralen Prinzipals an der Lösung eines eindeutig definierten Forschungsproblems (z.B. der Entwicklung eines medizinischen Wirkstoffs) arbeiten. Die Agenten wählen jeweils ein (gegebenenfalls für den Prinzipal nicht beobachtbares) kontinuierliches Anstrengungsniveau und zusätzlich je eine von zwei möglichen Forschungstechnologien. Beide Agenten können individuell das Forschungsproblem entweder erfolgreich lösen oder aber scheitern. Der individuelle und beobachtbare Forschungserfolg jedes Agenten steigt mit dem gewählten Anstrengungsniveau, jedoch nur dann, wenn der Agent eine *geeignete* Forschungstechnologie auswählt. Beide Technologien können unabhängig voneinander jeweils geeignet oder ungeeignet sein und im Falle der Wahl einer ungeeigneten Technologie scheitert die Forschung in jedem Fall, d.h. unabhängig von der erbrachten Anstrengung. Ex ante besteht über die Eignung beider Forschungstechnologien Unsicherheit, wobei die erste der beiden Technologien, die Mainstream-Technologie, im allgemeinen als wahrscheinlicher geeignet erachtet wird als die zweite (Outsider-Technologie).
Analysiert und gegenübergestellt werden drei unterschiedliche Informationsstrukturen im Zusammenspiel zwischen Prinzipal und Agenten. Im ersten Fall ist die Wahl von Anstrengung und Forschungstechnologie für den Prinzipal unmittelbar beobachtbar, sodass der

optimale Vertrag die Entlohnung direkt auf diese beide Variablen bedingt. Die Zuordnung der Agenten zu den jeweiligen Forschungstechnologien hängt von deren relativen Erfolgsaussichten ab und wird durch einen Schwellenwert definiert. Ist die Outsider-Technologie hinreichend wahrscheinlich geeignet, ist es aus Sicht des Prinzipals optimal, die Agenten aufzuteilen, sodass beide Agenten jeweils unterschiedliche Technologien einsetzen (diversifizierte Forschung). Andernfalls ist es optimal, beide Agenten mit der Mainstream-Technologie forschen zu lassen (konzentrierte Forschung).
Im zweiten Fall ist nur die Wahl der Technologie, nicht aber das Anstrengungsniveau der Agenten beobachtbar. Die Nicht-Beobachtbarkeit impliziert, dass der Prinzipal die Entlohnung der Agenten nur auf den Output (Erfolg bzw. Misserfolg) und nicht auf die Anstrengung selbst bedingen kann. Dies führt zu einem Effizienzverlust und resultiert zudem in einer Abnahme an diversifizierter Forschung, da wegen der Risikoaversion der Agenten die Wahl einer weniger aussichtsreichen Technologie mittels höherer Entlohnung kompensiert werden muss.
Im dritten Fall wird der optimale Vertrag zwischen Prinzipal und Agenten unter der Annahme, dass sowohl Anstrengungsniveau als auch die Wahl der Forschungstechnologie für den Prinzipal nicht beobachtbar sind, analysiert. Dies führt dazu, dass jeder Agent im Interesse der Maximierung seiner individuellen Erfolgsaussichten die Mainstream-Technologie der Outsider-Technologie vorzieht. Für den Fall, dass der Prinzipal diversifizierte Forschung bevorzugt, kommt es also zu einem Interessenkonflikt zwischen dem Prinzipal und demjenigen Agenten, der für die Forschung mit der Outsider-Technologie vorgesehen ist. Der optimale Vertrag berücksichtigt die Präferenz zur Mainstream-Technologie und bedingt die Entlohnung des für die Outsider-Technologie vorgesehenen Agenten auf seinen eigenen Output und den Output des anderen Agenten, da die Verteilung der Outputs Rückschlüsse über die Technologiewahl der Agenten zulässt und somit der ungewollte Einsatz der Mainstream-Technologie unterbunden werden kann. Die Anpassung des Vertrages geht mit einem weiteren Effizienzverlust einher und schränkt den Parameterraum, für welchen die diversifizierte Forschung optimal ist, abermals ein.

Der zweite Aufsatz beschäftigt sich erneut mit der optimalen Anreizgestaltung bei delegierter Forschung und analysiert das Problem strategisch motivierter Nicht-Offenlegung des tatsächlichen Fortschrittsniveaus eines Forschungsprojekts. In einem Modellrahmen, der demjeni-

gen des ersten Aufsatzes ähnelt, delegiert ein Prinzipal die Fertigstellung eines Forschungsprojekts an einen Agenten. Für den Abschluss des Vorhabens steht ein in zwei Perioden aufgeteilter Zeitraum zur Verfügung, sodass das Projekt entweder früh (in Periode 1), spät (in Periode 2) oder nie abgeschlossen werden kann. Hierbei wird angenommen, dass der Agent gegenüber dem Prinzipal über einen Informationsvorsprung verfügt, und dadurch die Preisgabe einer frühzeitigen Projektfertigstellung auf einen späteren Zeitpunkt verschieben kann. Für den Prinzipal ergibt sich dadurch ein erweitertes Problem optimaler Anreizgestaltung, da die Auszahlungen an den Agenten so gestaltet sein sollen, dass sie nicht nur die Erbringung optimaler Anstrengungsniveaus, sondern auch eine wahrheitsgemäße Offenlegung einer frühzeitigen Projektfertigstellung sicherstellen.
Das vorgestellte Modell analysiert den Parameterraum, für den ein Konflikt zwischen beiden Zielen existiert. Für den Fall, dass eine frühzeitige Fertigstellung des Projekts wünschenswert ist, ist es aus Sicht des Agenten nicht rational, dem Prinzipal diese vorzuenthalten. Der optimale anreizkompatible Vertrag stellt dann gleichzeitig die wahrheitsgemäße Offenlegung der Projektfertigstellung sicher. Wenn jedoch die Verhinderung eines ultimativen Scheiterns des Projekts das treibende Interesse des Prinzipals ist, existiert der genannte Zielkonflikt, und es kommt gegenüber der erstbesten Lösung zu Effizienzeinbußen.
Der Prinzipal reagiert auf den Anreiz des Agenten, den tatsächlichen Forschungsfortschritt zurückzuhalten mit der Anpassung des Vertrags. Eine erste mögliche optimale Anpassung besteht darin, die Auszahlungen an den Agenten derart anzupassen, dass diese ein Trenngleichgewicht induzieren und der Agent einen frühen Projektabschluss wahrheitsgemäß offenlegen wird. Die erwarteten Auszahlungen des Prinzipals sind der erstbesten Lösung (welche bei vollständiger Beobachtbarkeit des Projektfortschritts durch den Prinzipal erreichbar wäre) unterlegen, und es kommt zu einer Verzerrung der optimalen Anstrengungsniveaus in beiden Perioden. Somit ist bei der Wahl dieses Vertrags die strategische Verzögerung durch den Agenten zwar ein Problem, aber *kein* beobachtbares Phänomen auf dem Gleichgewichtspfad. Die zweite mögliche Anpassung des Vertrages besteht in einer Verschiebung des Projektstarts in die zweite Periode, sodass zur Fertigstellung des Projekts nur ein verkürzter Zeitraum zur Verfügung steht und der Prinzipal gegenüber der erstbesten Lösung ebenfalls schlechter gestellt ist. Die zweite Option wäre aus Sicht des Prinzipals dann optimal, wenn seine Auszahlung im Falle eines frühen Erfolgs (re-

lativ zur Auszahlung im Falle eines späten Erfolgs) hinreichend niedrig ist und wenn es dem Prinzipal möglich ist, den Agenten in der ersten Periode an der Durchführung der Forschung zu hindern.
Es werden weiterhin einige Modellerweiterungen und nicht vertragliche Lösungsansätze des Problems diskutiert. So wird gezeigt, dass etwa die Delegation des Projekts an zwei Agenten (anstatt wie zuvor an einen) das Ausmaß des Problems abschwächt. Zudem wird demonstriert, dass der Prinzipal durch die Überwachung des Agenten (sollte der Misserfolg eines Agenten durch eine mit Kosten verbundene Überprüfung verifizierbar sein) eine wahrheitswidriges Vorenthalten eines frühen Erfolgs unterbinden kann.

Der dritte Aufsatz beschäftigt sich mit dem Problem eingeschränkter bzw. kostspieliger Verifizierbarkeit von wissenschaftlichen Publikationen und und bietet eine spieltheoretische Analyse des akademischen Publikations- und Rezeptionsprozesses. Betrachtet wird das Zusammenspiel zwischen einem publizierenden Forschenden einerseits und einer an der Publikation interessierten relevanten wissenschaftlichen Community (der Leserschaft) andererseits. Die publizierten Resultate des Forschenden können entweder korrekt oder wissentlich inkorrekt (d.h. betrügerisch, z.B. durch Datenmanipulation etc.) sein. Jedes Mitglied der relevanten wissenschaftlichen Community hat die Möglichkeit, veröffentlichte Forschungsresultate unter Inkaufnahme privater Kosten zu überprüfen und, falls zutreffend, als falsch zu entlarven. Hierbei kommt es innerhalb der Leserschaft zu einem Trittbrettfahrerproblem, da der Nutzen der Überprüfung einer fehlerhaften Publikation der gesamten interessierten Wissenschaftsgemeinschaft als öffentliches Gut zugute kommt. Das Modell analysiert den Einfluss der Größe der jeweiligen wissenschaftlichen Community auf die Anreize des Forschenden, betrügerische Forschung zu veröffentlichen. Gezeigt wird die Existenz zweier symmetrischer Gleichgewichte, welche in Widerspruch zur Intuition stehen, dass Betrug mit zunehmender Größe der Leserschaft stärker abgeschreckt bzw. häufiger aufgedeckt wird. Im ersten Gleichgewicht bleibt das Ausmaß betrügerischer Artikel von einer wachsenden Leserschaft unbeeinflusst, wohingegen der Anteil der als betrügerisch aufgedeckten Publikationen ceteris paribus *sinkt*. Im zweiten Gleichgewicht stagniert bei größer werdender Leserschaft die Gesamtwahrscheinlichkeit, mit der eine betrügerische Publikation durch (mindestens) ein Mitglied der Leserschaft aufgedeckt wird, bei gleichzeitiger *Zunahme* des Umfangs betrügerischer Forschung.

Es wird zudem eine Reihe von Modellvarianten und -erweiterungen untersucht. Zunächst wird die mögliche unwissentliche Publikation inkorrekter Ergebnisse betrachtet, welche vom Forschenden durch ein hohes Sorgfaltsniveau vermieden werden kann. Hier zeigt sich ein potentiell negativer Einfluss der Community-Größe auf das Sorgfaltsniveau des Forschenden und einen dadurch bedingten höheren Anteil fehlerbehafteter Publikationen bei größeren Communitys. Zudem wird auch die Rolle der Zusammensetzung der wissenschaftlichen Leserschaft auf das Aufkommen (aufgedeckter) betrügerischer Forschung untersucht. Entgegen der Intuition sorgt ein höheres Maß an Diversität innerhalb der Leserschaft nicht notwendigerweise für einen geringeren Umfang an veröffentlichten betrügerischen Publikationen oder für ein höheres Ausmaß an aufgedecktem Forschungsbetrug. Weiterhin wird auch die Rolle der in der akademischen Welt vorherrschenden Prioritätsregel beleuchtet und gezeigt, dass diese die individuellen Anreize zur Überprüfung von veröffentlichter Forschung reduziert.
Forschende können zudem einen Anreiz haben, dass Trittbrettfahrerproblem innerhalb einer wissenschaftlichen Community absichtlich zu verstärken, indem die Größe der Leserschaft durch das Vortäuschen eindrucksvollerer Ergebnisse erhöht und somit die Wahrscheinlichkeit der eigenen Enttarnung verringert wird. Zuletzt wird gezeigt, dass auch der akademische Begutachtungsprozess adverse Effekte mit sich bringen kann, indem eine vor der Veröffentlichung vorgenommene Überprüfung der Publikation die Anreize für Forschende oder für die Leserschaft zur eigenen Überprüfung der Ergebnisse schmälert. In ihrer Gesamtheit legen die Ergebnisse nahe, dass der im Wissenschaftsbetrieb gängige Prozess aus Publikation und Verifikation weniger gut dazu geeignet ist, fehlerhafte Forschung zu verhindern oder als solche aufzudecken, als man bei unkritischer Betrachtung vermuten würde.

Summary of Contents

"*The scientist finds his reward in what Henri Poincaré calls the joy of comprehension, and not in the possibility of application to which any discovery may lead.*" (Albert Einstein, quoted in Planck (1932), p. 211). Many scientists nowadays would undoubtedly still agree with these remarks by Albert Einstein back in 1932. The joy of discovering new knowledge and understanding previously unknown relationships is probably one of the main driving forces for many researchers working inside and outside academia.

Even so, Einstein's image of a scientist driven solely by the urge to research and gain knowledge is an idealized one - at least from the perspective of economics. In the real world of academia, research funds and professorships are scarce, space in high-ranking academic journals is limited, and acknowledgement from colleagues is confined primarily to the handful among them who can come up with impressive results. From the vantage point of institutional economics, which emphasizes the importance of incentives, it can therefore be assumed that the pleasure of research per se cannot be the only motivation for scientists and that career concerns, for example, will also play a role. Viewed thus, the academic enterprise forms the institutional framework within which self-interested researchers adapt their behavior in such a way that their individual benefit is maximized. In this context, it is assumed - not very idealistically - that researchers will not shy away from opportunistic behavior. This can explain scientific misconduct, for example, or other questionable scientific practices such as "*p*-hacking", which find no place in Einstein's picture.

The scope for engaging in opportunistic behavior is particularly pronounced in research, as it is structurally characterized by information asymmetries. The research results achieved, the methods used and the research effort exerted are usually the private information of the respective researcher and can only be verified by third parties up to a point. It is therefore not surprising when Jean Tirole, for example, writes that the relationship between researchers and their funding sources is "fraught with moral hazard" (Tirole (2006)). At the same time, the work of researchers contributes significantly to technological progress and the emergence of product innovations, and is thus enormously important for economies to thrive and grow. Any frictions and misdirected incentives within science thus have a direct impact on overall welfare.

This dissertation therefore aims to forge a better understanding of incentives and institutions in (academic) research. It consists of three independent essays, each dealing with one specific problem area in the context of academic and industrial research from the perspective of institutional economics with the help of model-theoretical analyses.

The first paper addresses the individual selection of research approaches or research technologies by researchers in the context of *delegated* research. It shows that there might be an insufficient amount of diversity in the selection of research technologies used. A model is analyzed with two risk-averse agents (the researchers) working on behalf of a risk-neutral principal to solve a clearly defined research problem (e.g. the development of a new drug substance). The agents each choose a continuous level of effort (which may be unobservable to the principal) and, in addition, each choose one of two possible research technologies. Both agents can individually either successfully solve the research problem or fail. The individual and observable research success of each agent increases with the chosen level of effort, but only if the agent selects a *suitable* research technology. The two technologies can each be suitable or unsuitable independently of each other, and in the case where an unsuitable technology is chosen, the research fails in any case, i.e. no matter how much effort the researcher makes. Ex ante, there is uncertainty about the suitability of both research technologies, with the first of the two technologies, the mainstream technology, generally considered more likely to be suitable than the second (outsider technology).
Three different information structures in the interaction between principal and agent are analyzed and compared. In the first case, the choice of effort and research technology is directly observable to the principal, so that the optimal contract conditions the reward directly on these two variables. The assignment of agents to the respective research technologies depends on their relative likelihood of success and is defined by a threshold. If the outsider technology is sufficiently likely to be suitable, it is optimal from the principal's point of view to split the agents so that the agents each use different technologies (diversified research). Otherwise, it is optimal to have both agents research with the mainstream technology (concentrated research).
In the second case, only the choice of technology is observable, but not the effort level of the agents. This non-observability implies that the principal can condition the agents'

reward only on the output (success or failure) and not on the effort itself. This leads to a loss of efficiency and also results in a decrease in diversified research, since due to the agents' risk aversion the choice of a less promising technology has to be compensated for by a higher remuneration.
In the third case, the optimal contract between principal and agent is analyzed under the assumption that both the effort level and the choice of research technology are unobservable to the principal. This leads each agent to prefer the mainstream technology over the outsider technology in the interest of maximizing his individual chances of success. Thus, in the case where the principal prefers diversified research, a conflict of interest arises between the principal and the particular agent who is designated to conduct research with the outsider technology. The optimal contract takes into account the preference for the mainstream technology and conditions the reward of the agent designated to use the outsider technology on his own output and the output of the other agent, since the distribution of outputs allows conclusions to be drawn about the agents' choice of technology and thus the unwanted use of the mainstream technology can be prevented. The adjustment of the contract is accompanied by a further loss of efficiency and again restricts the parameter space for which diversified research is optimal.

The second paper revisits optimal incentive design in delegated research and analyzes the problem of strategically motivated non-disclosure of the actual progress level of a research project. In a model framework similar to that of the first paper, a principal delegates the completion of a research project to an agent. A two-period time frame is available for completion of the project, such that the project can be completed either early (in period 1), late (in period 2), or never. It is assumed that the agent has an information advantage over the principal and can therefore postpone the announcement of early project completion to a later point in time (strategic delay). For the principal, this results in an extended problem of optimal incentive design, since the payoffs to the agent should be designed to ensure not only the provision of optimal effort levels, but also truthful disclosure of early project completion.
The model presented analyzes the parameter space within which a conflict exists between the two goals. In the case where early completion of the project is desirable, it is not rational from the agent's point of view to withhold it from the principal. The

optimal incentive-compatible contract then simultaneously ensures truthful disclosure of the project's completion. However, if preventing the ultimate failure of the project is the principal's driving interest, the aforementioned trade-off exists, and efficiency losses occur relative to the first-best solution.

The principal responds to the agent's incentive to withhold actual research progress by adjusting the contract. A first possible optimal adjustment is to redesign the payoffs to the agent such that they induce a separating equilibrium and the agent will truthfully disclose early project completion. The principal's expected payoffs are inferior to the first-best solution (which would be achievable if the project's progress were fully observable to the principal), and there is a bias in the optimal effort levels in both periods. Thus, in the choice of this contract, strategic delay by the agent, while a problem, is *not* an observable phenomenon on the equilibrium path. The second possible adjustment to the contract is to shift the start of the project to the second period, leaving only a shortened period to complete the project and also leaving the principal worse off relative to the first-best solution. The second option would be optimal from the principal's point of view if her payoff in the case of early success (relative to the payoff in the case of late success) is sufficiently low and if it is possible for the principal to prevent the agent from performing the research in the first period.

Some model extensions and non-contractual remedies to the problem are further discussed. For example, it is shown that delegating the project to two agents (instead of one as before) mitigates the magnitude of the problem. In addition, it is demonstrated that by monitoring the agent (should an agent's failure be verifiable through costly verification), the principal can prevent an untruthful withholding of an early success.

The third paper deals with the problem of limited or costly verifiability of scientific publications and offers a game-theoretical analysis of the academic publication and reception process. The interaction between a publishing researcher on the one hand and a relevant scientific community (the readership) interested in the publication on the other hand is considered. The published results of the researcher can be either correct or knowingly incorrect (i.e. fraudulent, e.g. through data manipulation, etc.). Every member of the relevant scientific community has the possibility to verify published research results at their own personal expense and, if applicable, expose them as incorrect. Here, a free-rider

problem arises within the readership, since the benefit of reviewing an erroneous publication accrues to the entire interested scientific community as a public good. The model analyzes the influence of the size of the respective scientific community on the researcher's incentives to publish fraudulent research. Two symmetric equilibria are shown to exist that contradict the intuition that fraud is deterred more strongly or detected more frequently as the size of the readership increases. In the first equilibrium, the magnitude of fraudulent articles remains unaffected by increasing the size of the readership, whereas the proportion of publications detected as fraudulent decreases ceteris paribus. In the second equilibrium, as readership increases, the overall probability of a fraudulent publication being detected by (at least) one member of the readership stagnates, while the amount of fraudulent research *increases*.

A number of model variants and extensions are also examined. First, the possibility of unwittingly publishing incorrect results is considered, which can be avoided by the researcher by exercising a high level of due diligence. Community size is shown to have a potentially negative influence on the researcher's due diligence level, resulting in a higher proportion of erroneous publications in larger communities. In addition, the extent to which the composition of the scientific readership influences the occurrence of (detected) fraudulent research is also investigated. Contrary to intuition, a higher level of diversity within the readership does not necessarily ensure a lower level of published fraudulent publications or a higher level of uncovered research fraud. Further, the role of the priority principle prevalent in academia is also highlighted and shown to reduce individual incentives to review published research.

Researchers may also have an incentive to intentionally exacerbate the free-rider problem within a scientific community by increasing the size of the readership through the pretense of more impressive results, thereby reducing the likelihood that they will be exposed. Last, it is shown that the academic review process can also introduce adverse effects, in that a pre-publication review of the publication diminishes the incentives for researchers or for the readership to review the results themselves. Taken as a whole, the results suggest that the process of publication and verification commonly used in academia is less well suited to preventing or detecting flawed research than one might assume from an uncritical view.

Contents

List of Figures **XVIII**

List of Tables **XIX**

1 Contracting with Researchers **1**

1.1 Introduction . . . 2
1.2 Related Literature . . . 4
1.3 The Model . . . 6
1.3.1 Assumptions and Main Setting . . . 6
1.3.2 Contracting with a Single Researcher . . . 8
1.3.2.1 Symmetric Information . . . 8
1.3.2.2 Asymmetric Information . . . 9
1.3.3 Contracting with Two Researchers . . . 10
1.3.3.1 Symmetric Information . . . 11
1.3.3.2 Asymmetric Information . . . 16
1.4 Discussion . . . 22
1.5 Conclusion . . . 23

A Appendices **25**

A.1 Details on the Optimization Problems . . . 25
A.2 Proofs . . . 37

2 Strategic Delay in R&D Projects - An Agency Perspective **46**

2.1 Introduction . . . 47
2.2 Related Literature . . . 50
2.3 The Model . . . 52
2.3.1 First Best . . . 54
2.3.2 Delegated Research . . . 56
2.3.3 Optimal Contracting Under Strategic Delay . . . 58
2.4 Model extensions . . . 62
2.4.1 Limits to Enforceability . . . 62
2.4.2 Multiple Agents . . . 63

2.4.3 Replacement of Agents 65
2.4.4 Monitoring 65
2.4.5 Plurality of Research Methods 66
2.5 Discussion and Conclusion 67

B Appendices **68**
B.1 Proofs 68
B.2 Multiple Agents - Full Exposition 77

3 The Inspection Game in Science **81**
3.1 Introduction 82
3.2 Related Literature 85
3.3 The Model 87
3.3.1 Fraudulent Research: The Deception Game 87
3.3.2 Erroneous Research: The Delusion Game 95
3.4 Extensions and Applications 98
3.4.1 The Priority Principle in Science 98
3.4.2 Heterogeneous Readers and Ideological Diversity 99
3.4.3 Strategic Audience Choice 102
3.4.4 Editors and Peer Review 105
3.5 Discussion 107
3.6 Conclusion 108

C Appendix **110**

References **124**

List of Figures

1.1 a) The principal's payoff under symmetric information, when choosing concentrated efforts (mm) and diversified efforts (mo). b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts. 15

1.2 a) The principal's payoff under symmetric information (dashed) and under Moral Hazard I (solid) when choosing concentrated efforts (mm), and diversified efforts (mo). b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts. In the Moral Hazard I setting, the set of parameters for which DE is optimal shrinks (DE SB1). 19

1.3 a) The principal's payoff under Moral Hazard I (dashed), and under Moral Hazard II (solid), when choosing concentrated efforts (mm) and diversified efforts (mo). b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts. In the Moral Hazard II setting, the set of parameters for which DE is optimal shrinks (DE SB2). 21

3.1 The deception game under the simplified assumption of $n = 1$ reader . . . 90

3.2 Resulting equilibria for different realizations of G and B'. Brighter shades refer to a higher overall probability of fraud detection, given that a publication has been released. 93

3.3 The delusion game under the simplified assumption of $n = 1$ reader 95

3.4 Resulting equilibria for different realizations of G and B'. Brighter shades refer to a higher overall probability of error detection. 98

List of Tables

2.1 The principal's payoffs as a function of outcomes and reports in $t = 1, 2$ . . 53

1 Contracting with Researchers*

Matthias Verbeck†
University of Marburg

Elisabeth Schulte
University of Marburg

Abstract

We study a model of delegated research. A researcher's success depends on their effort and their choice of research technology which is uncertain with respect to its quality. Researchers pursue individual, rather than overall success, which yields a preference for the most promising technology. We show that a mechanism that deters this *bias towards mainstream research* always entails an efficiency loss if researchers are risk-averse. Our results suggest that there is too little diversity in delegated research.

Keywords: Moral hazard, Hidden action, Incentives in teams, Delegated research, Academic organization, Diversity in research

JEL codes: D82, D83, D86

*We wish to thank the participants in the 2015 annual meeting of the Verein für Socialpolitik, the participants in the internal research seminar run by the University of Marburg's Economics department, and the participants in the 2015 MAGKS research seminar in Rauischholzhausen for their very helpful comments and suggestions.

†Corresponding author: matthias.verbeck@gmail.com

1.1 Introduction

The job of a researcher - both in academia and in industry - is unique in many ways. One of its most striking characteristics is the high degree of uncertainty any researcher faces when trying to answer a specific research question. For all their hard work, success is by no means certain. Max Weber was certainly right when he wrote more than a century ago:

"Yet it is a fact that no amount of such enthusiasm, however sincere and profound it may be, can compel a problem to yield scientific results." (Weber (1946) [1917], p. 135).

Effort on the part of the researcher is a necessary but not a sufficient condition for success. In fact, the *technology* the researcher uses to address the research question is another important determinant of the probability of finding an answer. If a researcher backs the wrong horse - i.e. they use a method or technology that leads them into a dead end - all their efforts will be in vain. Hence, the choice of research technology is a risky bet from any researcher's perspective. In our model of research activities, we therefore separate a researcher's actions into "effort choice" and "technology choice". Two motivating examples will illustrate this issue:

Example 1: Development of vaccines
A national health agency is interested in the development of a vaccine for a fatal viral desease. Academic researchers or research teams can now choose from several methods to accomplish this goal. They could, for example, choose a conventional approach (e.g. development of a vector vaccine) or a novel approach (e.g. development of an RNA-based vaccine). If the research approach they choose turns out to be unsuitable, the researchers will stand no chance of success.

Example 2: Battery research
A car manufacturer wants to increase the range of the electrically powered vehicles it produces. Again, researchers have several options to choose from, e.g. they could try to improve the existing battery's cell chemistry, or they could pursue a completely new approach, e.g. developing a solid-state battery. Here too, only if the selected approach is

suitable will a higher level of effort increase the chances of success.
A second striking characteristic of research is its "winner-takes-all principle". The marginal value of a second successful researcher who replicates a discovery by one of their colleagues is (close to) zero. In the words of Dasgupta and Maskin:

"There is no value added when the same discovery is made a second, third, or fourth time. [...] [T]he winning research unit is the sole contributor to social surplus." (Dasgupta and Maskin (1987), p. 583).

Such "multiples" are likely to occur whenever researchers independently try to answer similar questions (Merton (1963)). From a social perspective, it is neither interesting how many researchers found the answer to a particular solved research question, nor which specific approach yielded the answer (abstracting from ethical considerations).[1] What *is* crucial is that there is an effective vaccine or a more powerful battery, say, not *how* it was found. From an ex ante perspective, however, in the face of uncertainty about any available research technology, a principal or social planner might find it optimal to diversify technological risk by pursuing more than one approach so as to maximize the overall probability of success. Our contract-theoretical analysis of this problem shows that this goal might conflict with the researchers' vested interests in maximizing their individual probability of success. Hence, whenever technology choice is not observable, a *moral hazard* problem arises because selfish researchers choose research portfolios that exhibit too little diversity from an efficiency perspective. In this paper, we study optimal contracting with researchers in the depicted setting.
Although diversity of research is often optimal from a social point of view, it is not rational from a researcher's perspective to make use of less promising technologies. Whether or not the assumption of an unobservable technology is reasonable, depends on the characteristics of the specific research question. For example, a non-expert principal might be sufficiently knowledgeable to be able to tell the difference between a vector vaccine and an RNA-based vaccine but lack the expertise needed to distinguish between individual lines of research within these different approaches. We will capture both cases (observable and

[1] If a definite answer to the research question has not yet been found, the number of researchers presenting a preliminary answer may still be informative. We focus on research questions for which only a definite answer is of value.

unobservable research technologies) in our analysis.

The remainder of the paper is organized as follows: Section 1.2 discusses related literature. Section 1.3 first analyzes the case of a single researcher. Here, the interests regarding the choice of technology are perfectly aligned. It then considers the option of contracting with a second researcher. We find that the optimal technology choice is distorted when the principal is unable to observe the selected technology. For such cases, the incentivization of the principal's preferred technology choice comes at a cost, such that there is an overall loss of efficiency. Section 1.4 discusses several critical assumptions and limitations of our model, and Section 1.5 concludes. Detailed derivations of the optimal contracts are collected in Appendix A.1. Detailed proofs can be found in Appendix A.2.

1.2 Related Literature

Our research is related to three branches of the economic literature. Methodologically, it contributes to the literature on incentives and incentives in teams. In the classic papers of Holmström ((1979), (1982)) the effort level is unobservable, leaving the principal with a lower expected return compared to the first-best solution. In our model, we extend the agent's strategy space and make the technology choice an (unobservable) part of any agent's strategy. Moreover, our research connects to Mookherjee (1984) and Itoh (1991), who both study compensation schemes for multiple agents and find that optimal individual remuneration should also depend on the other agents' performance in order to filter out common uncertainty. Legros and Matthews (1993) suggest a compensation scheme that deters free-riding by making use of different performance distributions of heterogeneous agents. Although free-riding is excluded from our model by our assumption of individually observable output, our model features similar characteristics, since differences in output distributions are harnessed to deter undesired actions by the agents. Hörner and Samuelson (2013) provide a model in which a single agent is incentivized to conduct costly experiments, but output - like in our model - also depends on the exogenously given project quality.

Second, our research contributes to the literature on the economics of research and development. Here, our results are linked to models that show an undue amount of aggregated research efforts in equilibrium, like Loury (1979) or Dasgupta and Stiglitz (1980). More specifically, our research is connected to models of optimal research portfolios, e.g. Bhat-

tacharya and Mookherjee (1986) or Dasgupta and Maskin (1987). The latter - like our model - shows that independent researchers choose research projects that are overly correlated from a social planner's perspective. In a model of Fershtman and Rubinstein (1997), two researchers independently conduct research at one of multiple sites to find a hidden treasure. In equilibrium, there is an efficiency loss due to a coordination failure, implying a wasteful duplication of research efforts. Moreover, more recent contributions to the theory of contests also show similarities to our model, e.g. the work of Erat and Krishnan (2012), and Konrad (2014). Our own contribution differs from the aforementioned models in a number of ways. First, and most importantly, the driving force for the wasteful duplication of research efforts in our model is the non-observability of research technologies. Moreover, and unlike in most of the models mentioned before, we explicitly assume that research is *delegated*, rather than independent. Hence, our model aims to capture research activities within a firm, instead of between (competing) firms. What is more, our model sheds light on how the prospects of different technologies explicitly influence the agent's optimal effort choice.

Third, in a wider context, our research is also related to the economics of science literature (Stephan (1996)), which analyzes the plethora of issues surrounding the creation and transfer of (academic) knowledge. Works cited here are only exemplary and incomplete. frey (2003), Starbuck (2005) and Grey (2010) criticize the prevailing system of academic peer-reviewed publication as unreliable, opaque, and discouraging to innovative research. Ioannidis (2005) identifies a publication bias towards false results. Kieser (2010) criticizes performance-related pay in academic research. Felgenhauer and Schulte (2014) show that an information loss between researcher and scientific community is implied when the researcher strives to publish their research.

None of the contributions we know of, however, deal with the issue of duplicated research efforts from an agency perspective. Therefore, our research is novel to the best of our knowledge.

1.3 The Model

1.3.1 Assumptions and Main Setting

A risk-neutral (female) principal is interested in a conclusive research outcome from a specified research question.[2] Her utility from a research project is given by

$$V(q, w) = q - E[W] \tag{1.1}$$

where $q \in \{0, 1\}$ denotes the stochastic output of the research project, which can be either a failure or a success. $W = \sum_{i=1}^{n} w_i$ denotes the overall compensation of the agent(s) employed, w_i denotes agent i's private wage, and n (the number of agents) is either one or two. Each agent employed chooses exactly one research technology $j \in \{m, o\}$, where technology m is labelled "mainstream technology"and technology o "outsider technology". Furthermore, each agent chooses a costly research effort level, $e_i \in \mathbb{R}_0^+$, that determines the probability of individual success. An agent's strategy can therefore be fully described as $(e_i, j) \in \mathbb{R}_0^+ \times \{m, o\}$ and his overall utility equals

$$U_i = u_i(w_i) - e_i. \tag{1.2}$$

As is standard in the literature, we assume that

$$u'(\cdot) > 0, \quad u''(\cdot) \leq 0.$$

The agents' reservation utility level is zero.

Any agent's individual output will depend on the choice of research technology, which can either be "good"or "bad", denoted by $\omega_j \in \{g, b\}$. We let $\pi_j = P(\omega_j = g)$ denote the probability that technology j is good and make the assumption that technologies are *independent*, i.e. knowing the quality of technology m is not informative about the quality of technology o.[3] We assume that $\pi_m \geq \pi_o$, i.e. the mainstream technology appears at least as promising as the outsider technology.

[2] You can imagine the principal as a social planner who wishes to maximize society's benefits from research, but she could just as well be a firm owner who wants to maximize gains from its R&D department.

[3] The independence assumption restricts the generality of our model, but simplifies the analysis. It is reasonable when technologies are sufficiently distinct from each other.

Let q_i denote the event that agent i's research yields a success. We impose:

$$P(q_i = 1 \mid e_i \times j) = \begin{cases} \rho(e_i), \text{ if } \omega_j = g \\ 0, \text{ else.} \end{cases} \tag{1.3}$$

A success is only possible if the agent has chosen a good technology; otherwise, all his efforts will be in vain. But even if a good technology has been chosen, success is not guaranteed and depends on the agent's effort. Here, $\rho(e_i)$ defines the probability of agent i's success, given that technology j is a good technology. As is standard in the literature, we assume that

$$\rho'(\cdot) > 0, \quad \rho''(\cdot) \leq 0.$$

We add the following technical assumptions that guarantee an interior solution:

$$\rho'(0) \cdot \pi_j > \tfrac{1}{u'(u^{-1}(0))}, \quad \rho(0) = 0, \quad \rho(\infty) = 1.$$

An agent's overall probability of success, assuming the use of technology j and effort e_i, is given as

$$P(q_i = 1 \mid e_i \times j) = \rho(e_i) \cdot \pi_j. \tag{1.4}$$

The principal offers the agent(s) a contract with a view to maximizing (1.1), anticipating that agent i will choose his actions so as to maximize (1.2). We assume that all of the above (number of agents, cost functions, utility functions, state probabilities) is common knowledge.

The course of action is as follows:

1. Nature chooses ω_j according to π_j.
2. The principal offers the agent(s) a take it or leave it contract to the agent(s) which they can either accept or reject.
3. If an agent accepts the contract, he chooses a technology and an effort level that maximizes his utility given the conditions of the contract. If the contract is rejected, the game ends.
4. Nature draws q_i according to (1.3) and each party is paid remuneration according to the specified conditions.

1.3.2 Contracting with a Single Researcher

In this section, we derive the optimal contract with a single researcher, $n = 1$. As there is no ambiguity, there is no need to use subscripts to denote individual agents. Let $\overline{w}$ indicate the wage level paid to the agent if his research yields a success and $\underline{w}$ the wage level if that research is fruitless.

1.3.2.1 Symmetric Information

As a benchmark, we start with the case of symmetric information. Postponing the choice of j, we obtain the following maximization problem:

$$\begin{aligned} E(V_j(\cdot)) = \underset{j,e,\underline{w},\overline{w}}{max}\, \pi_j \cdot [\rho(e) \cdot (1 - \overline{w}) + (1 - \rho(e)) \cdot (-\underline{w})] \\ +(1 - \pi_j) \cdot (-\underline{w}), \end{aligned} \tag{1.5}$$

subject to

$$\begin{aligned} \pi_j \cdot [\rho(e) \cdot u(\overline{w}) + (1 - \rho(e)) \cdot u(\underline{w})] + \\ (1 - \pi_j) \cdot u(\underline{w}) - e \geq 0. \end{aligned} \tag{1.6}$$

Using the Lagrangian to solve the principal's problem, we obtain the optimal co-insurance conditions (Borch (1962)) between principal and agent which yield

$$\overline{w} = \underline{w} = w. \tag{1.7}$$

The optimal effort and wage levels for a given technology are implicitly defined by

$$\rho'(e) \cdot \pi_j = \frac{1}{u'(w)} \tag{1.8}$$

and

$$w = u^{-1}(e). \tag{1.9}$$

The left-hand side of (1.8) equals the marginal product of effort and the right-hand side equals the marginal cost of effort, both seen from the principal's perspective. It is evident (and in accordance with intuition) that the optimal effort e and optimal wage w rise in π_j. As usual, under symmetric information, the risk-neutral principal completely insures the risk-averse agent against the risk of failure by paying a wage that is not conditioned on the agent's success.

Regarding the technology choice, we obtain the intuitive result that choosing the more

promising technology is optimal from the principal's perspective:

Proposition 1.1. *For $n = 1$ and symmetric information, technology m is the principal's optimal technology choice.*

Proof: Suppose it is true that the principal's payoff is higher when the agent chooses the outsider technology, so that $\rho(e) \cdot \pi_o - w(e,j) > \rho(e) \cdot \pi_m - w(e,j) \Leftrightarrow \pi_o > \pi_m$. This contradicts our assumption that $\pi_m \geq \pi_o$. □

As the principal can perfectly observe the agent's actions, she could induce the optimal choice of technology by inflicting (arbitrary) punishments on the agent for using the wrong technology. In equilibrium, punishing the agent for choosing the wrong technology is never necessary, as a rational agent does not benefit from diverging from the principal's optimal technology choice. With the possibility of a punishment, the agent is no longer indifferent between technologies and will certainly act in the best interest of the principal. The optimal contract is therefore written as $w(e,j)$.

1.3.2.2 Asymmetric Information

In the case of unobservable actions, the agent's incentive constraint becomes a part of the principal's optimization problem:

$$\begin{aligned} e_j \in argmax_{\hat{e}} \;\; \pi_j \cdot [\rho(\hat{e}) \cdot u(\overline{w}) + (1 - \rho(\hat{e})) \cdot u(\underline{w})] + \\ (1 - \pi_j) \cdot u(\underline{w}) - \hat{e}. \end{aligned} \tag{1.10}$$

We apply the common first-order approach (for details see Appendix A.1).
In order to satisfy the agent's participation constraint and the incentive compatibility constraint, it must be the case that

$$\overline{w} > \underline{w}. \tag{1.11}$$

We therefore obtain the typical result that success is rewarded and fruitless effort ($q = 0$) is punished. As usual, we see that the non-contractability of e entails an efficiency loss since

$$\begin{aligned} \pi_j \cdot \rho(e) \cdot u(\overline{w}) + (1 - \pi_j \cdot \rho(e)) \cdot u(\underline{w}) = e < \\ u(\pi_j \cdot \rho(e) \cdot \overline{w} + (1 - \pi_j \cdot \rho(e)) \cdot \underline{w}) \\ \Leftrightarrow u^{-1}(e) < \pi_j \cdot \rho(e) \cdot \overline{w} + (1 - \pi_j \cdot \rho(e)) \cdot \underline{w}, \end{aligned} \tag{1.12}$$

where the right-hand side is due to Jensen's inequality. The expected wage needed to induce a given effort level is thus larger under asymmetric information than under symmetric information.

Next, we extend the degree of asymmetric information and assume that the agent's technology choice is also not observable (or verifiable in court) by the principal. We will refer to the case of unobservable effort and observable technology choice as "Moral Hazard I" and unobservable effort and unobservable technology choice as "Moral Hazard II". The assumption of unobservable technology choice is - at least for many research settings - plausible, as any (non-expert) principal will find it difficult to observe the techniques and methods used by the agent(s).

Recall from Proposition 1.1 that it is optimal to choose the mainstream technology, from the principal's standpoint. It is easy to see that a rational agent will choose the same technology.

Proposition 1.2. *For n=1 and asymmetric information, technology m is the agent's optimal technology choice.*

Proof: Consider any (possibly suboptimal) effort choice by the agent. The agent's expected gain, when the mainstream technology is used, equals $\rho(e)\cdot\pi_m\cdot u(\overline{w})+(1-\rho(e)\cdot\pi_m)\cdot u(\underline{w})-e$. If we replace π_m with π_o, the expected gains are strictly lower, as $u(\overline{w})>u(\underline{w})$ and the higher utility level $u(\overline{w})$ gains a lower weight. Hence, the agent prefers technology m for any effort choice. □

The agent's and the principal's interests regarding technology choice are completely aligned, and the optimal contract only conditions on the output level, i.e. $w(q)$.

1.3.3 Contracting with Two Researchers

We now turn to the case where the principal can employ a second agent. The structure is similar to the one with a single agent. Each agent is assigned a specific technology and exerts research effort. The principal can choose to employ both agents, who either both use the same technology or use different technologies. We will refer to the former option as "concentrated efforts" and the latter as "diversified efforts". We make the important assumption that the *individual output* of agents is always observable, effectively excluding

free-riding problems from our setting. In addition, we assume that contracting *between agents* is impossible (no side-contracting) and that the researchers' probabilities of success are *independent* of each other and, thus, only depend on each researcher's own effort level and technology choice.

We are explicitly interested in cases where it is optimal for *both researchers* to exert a positive level of effort. This condition is not trivially satisfied, and there do indeed exist specific functions satisfying our previously defined specifications that yield an interior solution for one agent, but not for two. This is generally the case when an overall probability of success can be obtained in a cost-minimizing manner by leaving one agent completely inactive. Therefore, for $n = 2$, we add the following technical condition to our set of assumptions:

$$\pi_m \cdot \rho'(0) \cdot (1 - \rho(e^*)) > \frac{1}{u'(u^{-1}(0))} \tag{1.13}$$

where e^* labels the optimal effort level in the case $n = 1$. This condition is necessary to yield an interior solution for both agents which implies that employing two agents must result in a higher expected return than employing just one. For more details, see Appendix A.1.

In what follows, $\overline{\overline{w}}_i$ ($\overline{w}_i$) denotes agent i's wage level when both agents (only agent i) have been successful, and $\underline{\underline{w}}_i$ ($\underline{w}_i$) stands for the wage level when both agents (only agent i) fail.

1.3.3.1 Symmetric Information

Case 1 - Concentrated Efforts

We postpone the optimal choice of technologies and take it as given for determining the optimal effort and wage levels for each agent. With two agents, both using technology j, the principal's maximization problem equals

$$E(V_{jj}(\cdot)) = \max_{e_1,e_2,j,\overline{\overline{w}}_1,\overline{\overline{w}}_2,\overline{w}_1,\overline{w}_2,\underline{w}_1,\underline{w}_2,\underline{\underline{w}}_1,\underline{\underline{w}}_2} =$$

$$\begin{aligned}\pi_j \cdot [&\rho(e_1) \cdot \rho(e_2) \cdot (1 - \overline{\overline{w}}_1 - \overline{\overline{w}}_2) + \\ &\rho(e_1) \cdot (1 - \rho(e_2)) \cdot (1 - \overline{w}_1 - \underline{w}_2) + \\ &(1 - \rho(e_1)) \cdot \rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + \\ &(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\ &(1 - \pi_j) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2).\end{aligned} \tag{P II: FB CE}$$

The problem is subject to the individual rationality constraints of the agents (presented here only for agent 1, by analogy for agent 2):

$$\begin{aligned}\pi_j \cdot [&\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ &\rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\ &(1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\ &(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ &(1 - \pi_j) \cdot u(\underline{\underline{w}}_1) - e_1 \geq 0.\end{aligned} \tag{IR II: FB CE}$$

For both settings, concentrated efforts and diversified efforts, we once more obtain the result that the agents' wage only conditions on effort and not on performance:

$$\overline{\overline{w}}_i = \overline{w}_i = \underline{w}_i = \underline{\underline{w}}_i = w_i. \tag{1.14}$$

Plugging the uniform wage into the optimization problem, we can solve for optimal effort and wage levels. In Appendix A.2 we show that the necessary and sufficient conditions for positive effort levels for both agents are strictly satisfied.
We yield optimal effort and wage levels for agent 1 (likewise for agent 2) by solving

$$\rho'(e_1) \cdot \pi_j \cdot (1 - \rho(e_2)) = \frac{1}{u'(w_1)} \tag{1.15}$$

and

$$w_i = u^{-1}(e_i). \tag{1.16}$$

As can be seen from equation (1.15), optimal effort also depends on the probability of success of the other agent. We can show that identical effort levels for both agents are optimal.

Proposition 1.3. *Symmetric effort, i.e.* $e_1 = e_2 = e_i$*, is optimal when two agents use*

the same technology.

Proof: See Appendix A.2.

The intuition of the proof is that, given that condition (1.13) is met and it is optimal to employ two agents, any overall probability of success can be achieved in a cost-minimizing manner when both agents exert a symmetric level of effort.
Plugging the result stated in Proposition 1.3 into (1.15), optimal effort and wage levels are defined by

$$\rho'(e_i) \cdot \pi_j \cdot (1 - \rho(e_i)) = \frac{1}{u'(w_i)} \tag{1.17}$$

and (1.16). We can easily see that in the two-agent case a lower effort level per agent is optimal, since the left-hand side of (1.17) is smaller than the left-hand side of (1.8).

Case 2 - Diversified Efforts
If the agents each use one technology (agent 1 uses m and agent 2 uses o), the principal's problem becomes

$$E(V_{mo}(\cdot)) = \max_{e_1,e_2,j,\overline{\overline{w}}_1,\overline{\overline{w}}_2,\overline{w}_1,\overline{w}_2,\underline{w}_1,\underline{w}_2,\underline{\underline{w}}_1,\underline{\underline{w}}_2} =$$

$$\begin{aligned}
&(\pi_m \cdot \pi_o) \cdot [\rho(e_1) \cdot \rho(e_2) \cdot (1 - \overline{\overline{w}}_1 - \overline{\overline{w}}_2) + \\
&\rho(e_1) \cdot (1 - \rho(e_2)) \cdot (1 - \overline{w}_1 - \underline{w}_2) + \\
&(1 - \rho(e_1)) \cdot \rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + \\
&(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
&(\pi_m \cdot (1 - \pi_o)) \cdot [\rho(e_1) \cdot (1 - \overline{w}_1 - \underline{w}_2) + \\
&(1 - \rho(e_1)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
&((1 - \pi_m) \cdot \pi_o) \cdot [\rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + \\
&(1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
&((1 - \pi_m) \cdot (1 - \pi_o)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2),
\end{aligned} \tag{P II: FB DE}$$

subject to agent 1's participation constraint (likewise for agent 2)

$$\begin{aligned}(\pi_m \cdot \pi_o) \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ \rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\ (1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\ (1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ (\pi_m \cdot (1 - \pi_o)) \cdot [\rho(e_1) \cdot u(\overline{w}_1) + (1 - \rho(e_1)) \cdot u(\underline{\underline{w}}_1)] + \\ ((1 - \pi_m) \cdot \pi_o) \cdot [\rho(e_2) \cdot u(\underline{w}_1) + (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ ((1 - \pi_m) \cdot (1 - \pi_o)) \cdot u(\underline{\underline{w}}_1) - e_1 \geq 0.\end{aligned} \qquad \text{(IR II: FB DE)}$$

Plugging the uniform wage into the optimization problem, the optimal effort and wage levels solve

$$\pi_m \cdot \rho'(e_1) \cdot (1 - \pi_o \cdot \rho(e_2)) = \frac{1}{u'(w_1)} \tag{1.18}$$

for agent 1 and

$$\pi_o \cdot \rho'(e_2) \cdot (1 - \pi_m \cdot \rho(e_1)) = \frac{1}{u'(w_2)} \tag{1.19}$$

for agent 2, and (1.16) for both agents. From the previous two equations, we can conclude that agent 1, who uses the mainstream technology, exerts a higher level of effort than agent 2, who uses the outsider technology.[4]

Having derived optimal effort-wage combinations for diversified and concentrated efforts, we can now turn to the question of which of the two options is optimal. The principal has three possible options:

1. Both agents are assigned technology m (concentrated efforts I).
2. Both agents are assigned technology o (concentrated efforts II).
3. Agents are assigned alternate technologies (diversified efforts).

Following the reasoning of Proposition 1.1, assigning both agents the inferior technology cannot be optimal. Consequently, only the two remaining alternatives (concentrated efforts while using the mainstream technology and diversified efforts) have to be compared

[4]For the edge case of $\pi_m = \pi_o$, the effort levels would be identical.

to determine the optimal strategy. Diversified efforts are optimal when

$$E(V_{mo}(\cdot)) > E(V_{mm}(\cdot)), \tag{1.20}$$

where V_{mo} (V_{mm}) denotes the principal's payoff function for diversified efforts (concentrated efforts). We can show that there is a set of combinations of π_m and π_o where it is optimal to assign one agent the outsider technology. Letting e_1, w_i and $e_1^{'}, w_i^{'}$ denote the optimal effort and wage levels for concentrated and diversified efforts respectively, we obtain the following proposition:

Proposition 1.4. *For $n = 2$ with symmetric information, a threshold $\tilde{\pi}_o = (\pi_m \cdot (\rho(e_1) \cdot (2 - \rho(e_1)) - \rho(e_1')) - 2 \cdot w_1 + w_1' + w_2') / (\rho(e_2') \cdot (1 - \pi_m \cdot \rho(e_1'))) < \pi_m$ determines the optimal allocation of agents, where $\tilde{\pi}_o < \pi_m$ whenever $0 < \pi_m < 1$. When $\pi_o \leq \tilde{\pi}_o$, concentrated efforts with technology m are optimal; otherwise, diversified efforts are optimal.*

Proof: See Appendix A.2.

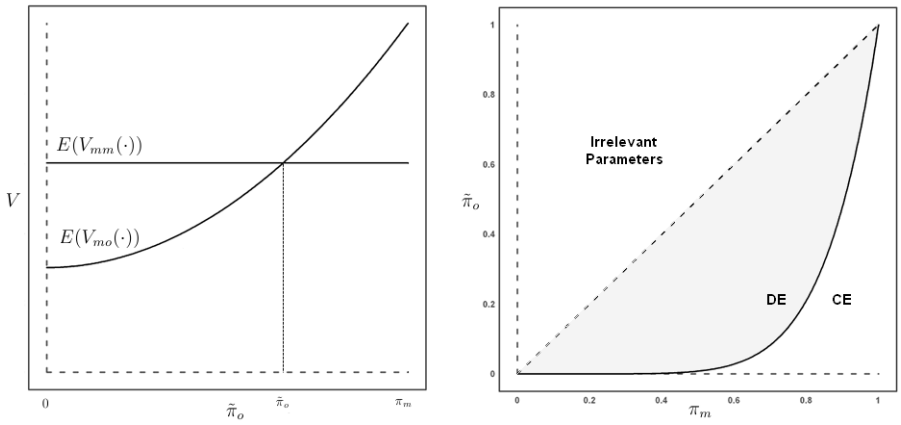

Figure 1.1:
a) The principal's payoff under symmetric information, when choosing concentrated efforts (mm) and diversified efforts (mo).
b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts.

The intuition for the proof is as follows: We know from Proposition 1.3 that if $\pi_o = 0$, concentrated efforts are better than diversified efforts. Next, we verify that for $\pi_o = \pi_m$ diversified efforts are strictly better than concentrated efforts. Finally, the fact that the principal's payoff from diversifying effort strictly increases in π_o, whereas her payoff from concentrated efforts is constant, yields a unique value for $\pi_o \in (0, \pi_m)$ where her payoffs are equal. Figure 1.1a illustrates the proof of Proposition 1.4 by showing the necessity of a unique intersection of $E(V_{mo}(\cdot))$ and $E(V_{mm}(\cdot))$ as a function of π_o. Figure 1.1b illustrates the optimal research portfolio for different parameter constellations of π_m and π_o, where "DE" and "CE" represent "diversified efforts" and "concentrated efforts", respectively. Note, that implementation of the optimal research agenda is again no problem in the symmetric information setting, since the uniform wage makes any agent indifferent between both technologies. Therefore, like in the case of a single agent, wrong technology choice can be prevented by an (arbitrarily small) punishment and the optimal contract is formulated as $w_i(e_i, j_i)$ for both agents.

1.3.3.2 Asymmetric Information

Let us now assume that the principal cannot observe the agents' actions, i.e, effort level and technology choice. We start with the analysis of observable technology choice and unobservable effort (Moral Hazard I).

Case 1 - Concentrated Efforts

When effort is unobservable, we have to add incentive compatibility constraints to the principal's maximization problem when contracting with agent 1 and agent 2, respectively (see Appendix A.1). Applying the first-order approach and then constructing the Lagrangian yields

$$\frac{1}{u'(\overline{\overline{w}}_i)} = \frac{1}{u'(\overline{w}_i)} = \lambda_i + \mu_i \cdot \frac{\rho'(e_i)}{\rho(e_i)}, \tag{1.21}$$

and

$$\frac{1}{u'(\underline{w}_i)} = \lambda_i - \mu_i \cdot \frac{\rho'(e_i)}{1-\rho(e_i)}, \tag{1.22}$$

and, by taking into account that $e_1 = e_2 = e_i$,

$$\frac{1}{u'(\underline{\underline{w}}_i)} = \lambda_i - \mu_i \cdot \frac{\pi_m \cdot \rho'(e_i) \cdot (1-\rho(e_i))}{\pi_m \cdot (1-\rho(e_i))^2 + (1-\pi_m)}, \tag{1.23}$$

where λ_i and μ_i are Lagrange multipliers.

From equations (1.21) to (1.23), we can derive the structure of optimal wages for concentrated efforts. Letting $\overline{\overline{w}}_i^{SB1}, \overline{w}_i^{SB1}, \underline{w}_i^{SB1}$, and $\underline{\underline{w}}_i^{SB1}$ denote the optimal wages[5] for agent i for different output distributions (analogous to the case of symmetric information), we obtain the following result:

Lemma 1.1. *For $n = 2$ with unobservable effort and observable technology choice, the structure of optimal wages when efforts are concentrated is $\overline{\overline{w}}_i^{SB1} = \overline{w}_i^{SB1} > \underline{\underline{w}}_i^{SB1} \geq \underline{w}_i^{SB1}$.*

Proof: $\overline{\overline{w}}_i^{SB1} = \overline{w}_i^{SB1}$ follows directly from (1.21). $\overline{\overline{w}}_i^{SB1} > \underline{\underline{w}}_i^{SB1}$ must be true, because the right-hand side of equation (1.21) is strictly larger than the right-hand side of equation (1.23). Furthermore, $\underline{\underline{w}}_i^{SB1} \geq \underline{w}_i^{SB1}$ follows from comparing (1.22) to (1.23), where the inequalities are identical except for the term $(1 - \pi_m)$ that is added to the right-hand side denominator of (1.23), such that we have a strict inequality whenever $\pi_m < 1$. □

If an agent fails to produce a positive output, his wage also depends on the performance of the other agent. This is the case because the other agent's output is informative about the technology's quality, and individual performance alone is not a sufficient statistic for any agent's effort level (Mookherjee (1984)). By incorporating into the contract any signal that is informative with respect to individual effort choice (Holmström (1979)), a more advantageous trade-off between effort provision and insurance is created for the principal.

Case 2 - Diversified Efforts

As in the case of concentrated efforts, we have to add the agents' incentive constraints to the original problem. This yields

$$\frac{1}{u'(\overline{\overline{w}}_i)} = \frac{1}{u'(\overline{w}_i)} = \lambda_i + \mu_i \cdot \frac{\rho'(e_i)}{\rho(e_i)}, \tag{1.24}$$

$$\frac{1}{u'(\underline{w}_1)} = \frac{1}{u'(\underline{\underline{w}}_1)} = \lambda_1 - \mu_1 \cdot \frac{\pi_m \cdot \rho'(e_1)}{1 - \ \pi_m \cdot \rho(e_1)} \tag{1.25}$$

as well as

$$\frac{1}{u'(\underline{w}_2)} = \frac{1}{u'(\underline{\underline{w}}_2)} = \lambda_2 - \mu_2 \cdot \frac{\pi_o \cdot \rho'(e_2)}{1 - \ \pi_o \cdot \rho(e_2)}. \tag{1.26}$$

[5]The superscript refers to the second-best solution in the presence of Moral Hazard I.

From the previous equations we can derive the wage structure for diversified efforts:

Lemma 1.2. *For $n = 2$ with unobservable effort and observable technology choice, the structure of optimal wages when efforts are diversified is $\overline{\overline{w}}_i'^{SB1} = \overline{w}_i'^{SB1} > \underline{w}_i'^{SB1} = \underline{\underline{w}}_i'^{SB1}$.*

Proof: $\overline{\overline{w}}_i'^{SB1} = \overline{w}_i'^{SB1}$ follows directly from (1.24). $\overline{w}_i'^{SB1} > \underline{w}_i'^{SB1}$ must be true, because the right-hand side of equations (1.25) and (1.26) is strictly larger than the right-hand side of equation (1.24). $\underline{w}_i'^{SB1} = \underline{\underline{w}}_i'^{SB1}$ follows directly from (1.25) and (1.26). □

Due to the technological independence, the performance of agent 1 is not a signal for the effort level of agent 2, and vice versa. Hence, when efforts are diversified, individual performance alone determines the wage level for any agent. This independence will massively facilitate our analysis of the problem.
In both cases, concentrated and diversified efforts, under asymmetric information, the expected wage for any agent needed to induce the first-best effort level, is higher than under symmetric information. The reasoning is similar to that in the single-agent setting, so we shall refrain from formally restating the argument here. What is more interesting is the change in the optimal research portfolio generated by the non-observability of effort. We obtain the intuitive result that concentrated efforts (i.e. research with the more promising technology) are optimal for more parameter constellations of π_m and π_o as compared to the first-best solution. Letting $E(W_i^{SB1})$ and $E(W_i'^{SB1})$ denote the expected wage level of agent i for concentrated and diversified efforts respectively, we obtain the following proposition:

Proposition 1.5. *For $n = 2$ and unobservable effort (Moral Hazard I,*
$\tilde{\pi}_o^{SB1} = \left(\pi_m \cdot \left(\rho\left(e_1^{SB1}\right)\right) \cdot \left(\left(2 - \rho(e_1^{SB1})\right) - \rho\left(e_1'^{SB1}\right)\right) - 2 \cdot E\left(W_1^{SB1}\right) + E\left(W_1'^{SB1}\right) + E\left(W_2'^{SB1}\right)\right) /$
$\left(\rho\left(e_2'^{SB1}\right) \cdot \left(1 - \pi_m \cdot \rho\left(e_1'^{SB1}\right)\right)\right)$ *determines the optimal allocation of agents, where $\tilde{\pi}_o < \tilde{\pi}_o^{SB1} < \pi_m$ whenever $0 < \pi_m < 1$. When $\pi_o \leq \tilde{\pi}_o^{SB1}$, concentrated efforts with technology m are optimal, otherwise diversified efforts are optimal.*

Proof: See Appendix A.2.

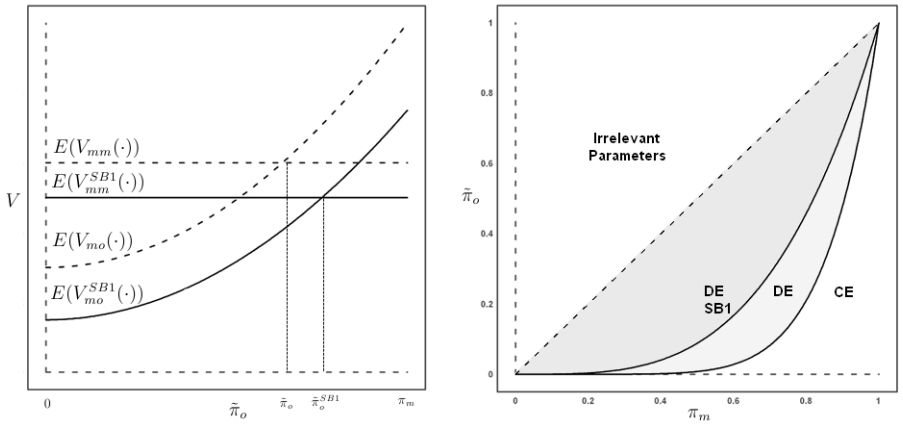

Figure 1.2:
a) The principal's payoff under symmetric information (dashed) and under Moral Hazard I (solid) when choosing concentrated efforts (mm), and diversified efforts (mo).
b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts. In the Moral Hazard I setting, the set of parameters for which DE is optimal shrinks (DE SB1).

The intuition for proof of the unique threshold's existence is similar to that proof of Proposition 1.4. Proof that $\tilde{\pi}_o^{SB1} > \tilde{\pi}_o$ is obtained by comparing the second agent's marginal gain from investing effort with technology m and o, respectively. From this it is possible to derive an upper bound for $\tilde{\pi}_o$ and a lower bound for $\tilde{\pi}_o^{SB1}$. Under asymmetric information, investing effort with technology o is disproportionally more expensive than investing effort with technology m. This is due to the increased technological risk which an agent has to bear when working with technology o. The higher relative cost of using technology o causes the set of parameter constellations of π_m and π_o for which concentrated efforts are optimal to increase. Figure 1.2 illustrates this argument. The optimal contractsfor agent 1 and agent 2 respectively, are written as $w_1(q_1, q_2)$ and $w_2(q_1, q_2, j_2)$. Note that it is only agent 2, whose contract refers to technology choice. This is because it is in both agents' interests to choose technology m, and agent 2 must be deterred, by means of a sufficiently large fine, from doing so (cf. Proposition 1.6).
Next we turn to the problem of Moral Hazard II, where technology choice is likewise unobservable. Following the reasoning of Proposition 1.2, any agent would prefer to use the mainstream technology. Hence, there is no need to explicitly incentivize agents to choose the mainstream technology, and asymmetric information with respect to technology choice does not harm the principal when the parameter constellation is such that

she prefers concentrated efforts. Whenever diversified efforts are optimal, however, the wage scheme derived for the previous problem is no longer optimal. In fact, under that wage scheme, agent 2 will deviate from the principal's desired behavior and switch to the mainstream technology, since choosing that technology increases the probability of individual success for any given effort level.

Proposition 1.6. *For n=2 with unobservable effort and unobservable technology choice (Moral Hazard II), every agent will choose the mainstream technology under the optimal wage scheme for Moral Hazard I.*

Proof: See Appendix A.2.

Since the agents' technology choice cannot be observed, there needs to be a mechanism that incentivizes agent 2 to optimally choose the outsider technology. Hence, we have to add another incentive compatibility constraint for the second agent:

$$\begin{aligned}
&\pi_m \cdot \pi_o \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \rho(e_1) \cdot (1-\rho(e_2)) \cdot u(\underline{w}_2) + \\
&(1-\rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
&(\pi_m \cdot (1-\pi_o)) \cdot [\rho(e_1) \cdot u(\underline{w}_2) + (1-\rho(e_1)) \cdot u(\underline{\underline{w}}_2)] + \\
&((1-\pi_m) \cdot \pi_o) \cdot [\rho(e_2) \cdot u(\overline{w}_2) + (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
&(1-\pi_m) \cdot (1-\pi_o) \cdot u(\underline{\underline{w}}_2) - e_2 \geq \\
&\pi_m \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \rho(e_1) \cdot (1-\rho(e_2)) \cdot u(\underline{w}_2) + \\
&(1-\rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
&(1-\pi_m) \cdot u(\underline{\underline{w}}_2) - e_2.
\end{aligned} \tag{IC2 II: SB2 DE}$$

Incorporating the additional constraint into the principal's problem yields a new wage scheme that rewards or punishes the second agent according to his own performance *and* the performance of agent 1.

Lemma 1.3. *For agent 2, the structure of wages for diversified unobservable efforts and unobservable technology choice is* $\overline{w}_2^{SB2} > \overline{\overline{w}}_2^{SB2}$ *and* $\underline{w}_2^{SB2} > \underline{\underline{w}}_2^{SB2}$.

Proof: See Appendix A.2.

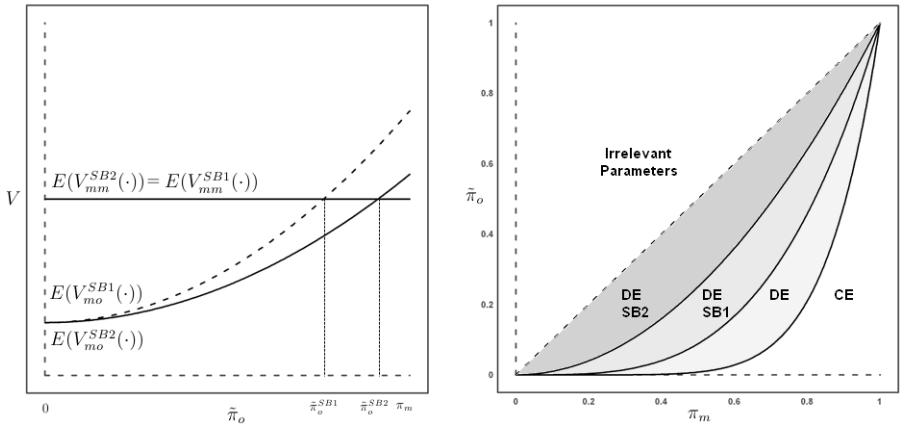

Figure 1.3:
a) The principal's payoff under Moral Hazard I (dashed), and under Moral Hazard II (solid), when choosing concentrated efforts (mm) and diversified efforts (mo).
b) The optimal research portfolio for different parameter constellations (π_m, π_o). CE: concentrated efforts, DE: diversified efforts. In the Moral Hazard II setting, the set of parameters for which DE is optimal shrinks (DE SB2).

The resulting wage scheme for agent 2 is a function of q_1 and q_2 and rewards him according to the distribution of outcomes.[6] If he is the sole agent to succeed, his earnings are higher compared to the outcome where both agents are successful. Moreover, his punishment is more severe when both agents fail compared to a situation where he alone fails. This new wage structure is optimal because output distributions are informative about agent 2's compliance with the principal's desired actions. A single success and a single failure are more likely to occur in situations where the agent complied with the principal's wishes. A double success and a double failure, however, are signals of deviant behavior. Our wage scheme is therefore similar to the one proposed by Legros and Matthews (1993), who use heterogeneity in agents' characteristics to deter free-riding in team production.
Incentivizing her preferred technology choice comes at a cost for the principal because of the agents' risk aversion. For both performance levels of agent 2, his respective wage also depends on the performance of agent 1. Hence, he faces a lottery in both cases. Agent 2 prefers pairwise sure outcomes over lotteries in each case with an identical expected value. Therefore, an additional risk premium is necessary to ensure the agent's participation, which entails a cost for the principal.

Proposition 1.7. *The principal's expected payoff is lower for diversified efforts, when the*

[6]The optimal contract for agent 1 still only depends on the outcome produced by agent 1.

technology choice is not observable: $E(V_{mo}^{SB1}(\cdot)) > E(V_{mo}^{SB2}(\cdot))$

Proof: See Appendix A.2.

Since diversified efforts are more costly compared to Moral Hazard I, the set of parameter constellations for which diversified efforts are optimal shrinks further, and we define a new threshold.

Proposition 1.8. *For* $n = 2$ *with unobservable effort and unobservable technology choice (Moral Hazard II),* $\tilde{\pi}_o^{SB2} = (\pi_m \cdot (\rho(e_1^{SB2}) \cdot (2 - \rho(e_1^{SB2})) - \rho(e_1'^{SB2})) - 2 \cdot E(W_1^{SB2}) + E(W_1'^{SB2}) + E(W_2'^{SB2}))/(\rho(e_2'^{SB2}) \cdot (1 - \pi_m \cdot \rho(e_1'^{SB2})))$ *determines the optimal allocation of agents, where* $\tilde{\pi}_o^{SB1} < \tilde{\pi}_o^{SB2} < \pi_m$ *whenever* $0 < \pi_m < 1$. *When* $\pi_o \leq \tilde{\pi}_o^{SB2}$, *concentrated efforts with technology* m *are optimal, otherwise diversified efforts are optimal.*

Proof: See Appendix A.2.

Again, we have a non-empty set of parameter constellations for which it is optimal to diversify, although this set must be smaller than under Moral Hazard I. Figure 1.3 illustrates Proposition 1.8.

1.4 Discussion

Our results suggest that - for the case of two researchers - the individually optimal research portfolio choice need not coincide with the social optimum when asymmetric information between the principal and her agents is involved. First, when effort is unobservable but technology choice is observable, the adjusted optimal research portfolio shifts towards the mainstream technology for a larger set of parameter realizations compared to the first-best solution (Moral Hazard I). Second, the wage scheme developed for Moral Hazard I is no longer optimal if the principal wants to induce a multiplicity of research approaches and technology choice is also unobservable for the principal. Absent modifications to the wage structure, only the mainstream technology would be used by both agents. The optimal wage scheme for Moral Hazard II takes into account that an agent who is supposed to use the inferior technology has to be additionally incentivized to do so. However, this adjustment to payments comes at a cost for the principal, shifting the optimal research portfolio once more towards mainstream research for more parameter constellations. Hence, the

bias towards mainstream technology becomes more pronounced in Moral Hazard II. Unlike in related models such as Dasgupta and Maskin (1987) and Fershtman and Rubinstein (1997), the misdirection of research effort is completely due to the information asymmetry between principal and agents.

Admittedly, our model is quite stylized and does not cover important aspects of reality. First and foremost, the resulting bias towards mainstream research is (partly) driven by the assumptions that all players have a common prior about success probabilities, identical cost functions for both technologies, and decide simultaneously which technology to chose. Changes in these assumptions might yield results that resolve or mitigate the resulting bias. Furthermore, we exclude important aspects like economies of scale[7] and closely related technologies, which breach our assumption of independence. One might also criticize the resulting optimal wage scheme for being too complicated or unrealistic. As valid as all these points may be, our model does help to understand why research diversity is not something that is easily achieved or follows naturally from a researcher's own interest. In fact, without well-designed incentives, a beneficial multiplicity of research approaches is not likely to occur.

1.5 Conclusion

We have derived optimal contracts for a setting of delegated research, in which the agents' action space encompasses an effort level and the choice between two research technologies. For a single agent, the optimal second-best contract is simple, and characterized by an effort level that is higher the more promising the superior technology. Optimal technology choice follows from the agent's self-interest and does not have to be incentivized by the contract. Hence, the non-observability of effort reduces the principal's expected income, whereas the non-observability of technology choice does not.

For two agents, depending on the respective realizations of parameter values, either (i) concentrating efforts on the mainstream technology or (ii) diversifying efforts on both technologies can be optimal. Given technological independence, the optimal second-best contract is conditioned on the other agent's performance level only when efforts are con-

[7] For example, success probabilities might disproportionally increase when more than one agent uses a certain technology, due to knowledge spillovers.

centrated. Unobservable effort shifts the optimal allocation of researchers towards the mainstream technology for a larger range of parameter values compared to the first-best solution. When the principal intends to induce diversified efforts and technology choice cannot be observed, the original second-best wage scheme fails, since using the mainstream technology will always yield the agent a higher expected payoff. The desired choice of technology can be induced by an adjusted payoff scheme that harnesses differences in outcome distributions. The distortion caused by the additional information asymmetry lowers the principal's expected payoff and leads to a further enlargement of the set of parameters for which concentrated efforts are optimal. Our model suggests that there is a socially suboptimal level of diversity in research when multiple researchers work on an identical research goal.

A Appendices

A.1 Details on the Optimization Problems

Symmetric Information, n=1

From (1.5) and (1.6) we obtain the Lagrangian

$$\mathcal{L} = \begin{array}{c} \rho(e) \cdot \pi_j \cdot (1 - \overline{w}) + (1 - \rho(e) \cdot \pi_j) \cdot (-\underline{w}) + \\ \lambda \cdot [\rho(e) \cdot \pi_j \cdot u(\overline{w}) + (1 - \rho(e) \cdot \pi_j) \cdot u(\underline{w}) - e] = 0. \end{array} \tag{A.1}$$

Taking the first-order conditions yields

$$\frac{\partial \mathcal{L}}{\partial e} = \begin{array}{c} \rho'(e) \cdot \pi_j \cdot (1 - \overline{w} + \underline{w}) + \\ \lambda \cdot [(\rho'(e) \cdot \pi_j \cdot (u(\overline{w}) - u(\underline{w})) - 1] = 0, \end{array} \tag{A.2}$$

$$\frac{\partial \mathcal{L}}{\partial \overline{w}} = \rho(e) \cdot \pi_j \cdot (-1) + \lambda \cdot [\rho(e) \cdot \pi_j \cdot u'(\overline{w})] = 0, \tag{A.3}$$

$$\frac{\partial \mathcal{L}}{\partial \underline{w}} = \begin{array}{c} (1 - \rho(e) \cdot \pi_j) \cdot (-1) + \\ \lambda \cdot [(1 - \rho(e) \cdot \pi_j) \cdot u'(\underline{w})] = 0. \end{array} \tag{A.4}$$

From (A.3) and (A.4) we can easily obtain the optimal co-insurance conditions and yield

$$\frac{1}{u'(\overline{w})} = \frac{1}{u'(\underline{w})} \Leftrightarrow u'(\overline{w}) = u'(\underline{w}) \Leftrightarrow \overline{w} = \underline{w}. \tag{A.5}$$

Plugging the uniform wage w into (A.2) we yield

$$\frac{1}{u'(w)} = \lambda = \frac{\rho'(e) \cdot \pi_j \cdot (1 - w + w)}{1 - \rho'(e) \cdot \pi_j \cdot (u(w) - u(w))} \tag{A.6}$$

which can be rearranged to (1.8).

Asymmetric Information, n=1

We obtain the following program

$$E(V_j^{SB}(\cdot)) = \underset{j,e,\underline{w},\overline{w}}{max}\, \pi_j \cdot [\rho(e) \cdot (1-\overline{w}) + (1-\rho(e)) \cdot (-\underline{w})] \\ +(1-\pi_j) \cdot (-\underline{w}), \tag{P I: SB}$$

subject to

$$\pi_j \cdot [\rho(e) \cdot u(\overline{w}) + (1-\rho(e)) \cdot u(\underline{w})] + \\ (1-\pi_j) \cdot u(\underline{w}) - e \geq 0 \tag{IR I: SB}$$

and

$$e_j \in argmax_{\hat{e}} \pi_j \cdot [\rho(\hat{e}) \cdot u(\overline{w}) + (1-\rho(\hat{e})) \cdot u(\underline{w})] + \\ (1-\pi_j) \cdot u(\underline{w}) - \hat{e}. \tag{IC I: SB}$$

To solve this problem, we can use the common first-order condition approach, given our assumptions on $\rho(\cdot)$, $u(\cdot)$, and π_j. Thus, the agent's original incentive constraint is replaced by

$$\rho'(e) \cdot \pi_j \cdot [u(\overline{w}) - u(\underline{w})] = 1. \tag{A.7}$$

We obtain the Lagrangian

$$\begin{aligned} \mathcal{L} = {} & \rho(e) \cdot \pi_j \cdot (1-\overline{w}) + (1-\rho(e) \cdot \pi_j) \cdot (-\underline{w}) + \\ & \lambda \cdot [\rho(e) \cdot \pi_j \cdot u(\overline{w}) + (1-\rho(e) \cdot \pi_j) \cdot u(\underline{w}) - e] + \\ & \mu \cdot [\rho'(e) \cdot \pi_j \cdot [u(\overline{w}) - u(\underline{w})] - 1] = 0. \end{aligned} \tag{A.8}$$

Taking derivatives with respect to $\overline{w}$ and $\underline{w}$ yields

$$\frac{1}{u'(\overline{w})} = \lambda + \mu \cdot \frac{\rho'(e)}{\rho(e)} \tag{A.9}$$

and

$$\frac{1}{u'(\underline{w})} = \lambda - \mu \cdot \frac{\rho'(e) \cdot \pi_j}{1-\rho(e) \cdot \pi_j}. \tag{A.10}$$

Equations (A.9) and (A.10) imply that $\overline{w} > \underline{w}$.

Symmetric Information, n=2

Concentrated Efforts

A necessary condition for $e_1, e_2 > 0$

To see that a single agent might be optimal in some cases, consider the following specific functions that satisfy the assumptions we made in section 1.3.1. $\rho(e_i) = \frac{e_i+{e_i}^2+{e_i}^3+{e_i}^4}{e_i+{e_i}^2+{e_i}^3+{e_i}^4+1}$ and $u(w_i) = \frac{ln(w_i+1)}{ln(2)}$. The latter function implies $u_i^{-1}(e_i) = 2^{e_i} - 1$. For a single agent, $\pi_m \cdot \rho'(0) > \frac{1}{u'(u^{-1}(0))} \Leftrightarrow \pi_m \cdot 1 > ln(2)$ is clearly fulfilled for sufficiently large levels of π_m, such that an interior solution must exist. For the particular case of $\pi_m = 1$, the optimal effort level can be approximated as $e^* \approx 0.41$, implying that $\rho(e^*) \approx 0.4$. Therefore, if agent 1 exerts the optimal effort level for a single agent, agent 2 should not invest any effort at all, since condition (1.13) is violated: $1 \cdot 1 \cdot (1 - 0.4) < ln(2)$. In fact, it can be shown that for the particular functions above, any combination of positive effort levels for both agents is inferior to the combination $e_1 = e^*, e_2 = 0$, so that a rational principal would always hire only a single agent. However, inequality (1.13) is not overly restrictive in general, and a large fraction of functions that meet the conditions from section 1.3.1 will also fulfil inequality (1.13).

Further details:

From the maximization problem we obtain the Lagrangian

$$\begin{aligned}
\mathcal{L} = \pi_j \cdot [&\rho(e_1) \cdot \rho(e_2) \cdot (1 - \overline{\overline{w}}_1 - \overline{\overline{w}}_2) + \\
&\rho(e_1) \cdot (1 - \rho(e_2)) \cdot (1 - \overline{w}_1 - \underline{w}_2) + \\
&(1 - \rho(e_1)) \cdot \rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + \\
&(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
&(1 - \pi_j) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)) + \\
\lambda_1 \cdot [\pi_j \cdot [&\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\
&\rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\
&(1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\
&(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\
&(1 - \pi_j) \cdot u(\underline{\underline{w}}_1) - e_1] + \\
\lambda_2 \cdot [\pi_j \cdot [&\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \\
&\rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\underline{w}_2) + \\
&(1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + \\
&(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
&(1 - \pi_j) \cdot u(\underline{\underline{w}}_2) - e_2] = 0.
\end{aligned} \tag{A.11}$$

Taking derivatives with respect to the different wage levels for agent 1 (likewise for agent 2) yields

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \overline{\overline{w}}_1} = &\pi_j \cdot \rho(e_1) \cdot \rho(e_2) \cdot (-1) + \\
&\lambda_1 \cdot [\pi_j \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] = 0 \\
&\Leftrightarrow \frac{1}{u'(\overline{\overline{w}}_1)} = \lambda_1,
\end{aligned} \tag{A.12}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \overline{w}_1} = &\pi_j \cdot \rho(e_1) \cdot (1 - \rho(e_2)) \cdot (-1) + \\
&\lambda_1 \cdot [\pi_j \cdot \rho(e_1) \cdot (1 - \rho(e_2)) \cdot u'(\overline{w}_1)] = 0 \\
&\Leftrightarrow \frac{1}{u'(\overline{w}_1)} = \lambda_1,
\end{aligned} \tag{A.13}$$

$$\begin{aligned}
&\pi_j \cdot (1-\rho(e_{1j})) \cdot \rho(e_2) \cdot (-1) + \\
\frac{\partial \mathcal{L}}{\partial \underline{w}_1} = \lambda_1 \cdot [\pi_j \cdot (1-\rho(e_1)) \cdot \rho(e_2) \cdot u'(\underline{w}_1)] &= 0 \\
\Leftrightarrow \frac{1}{u'(\underline{w}_1)} &= \lambda_1,
\end{aligned} \tag{A.14}$$

and

$$\begin{aligned}
&\pi_j \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot (-1) + \\
&(1-\pi_j) \cdot (-1) + \\
\frac{\partial \mathcal{L}}{\partial \underline{\underline{w}}_1} = \lambda_1 \cdot [\pi_j \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u'(\underline{\underline{w}}_1) + &\\
(1-\pi_j) \cdot u'(\underline{\underline{w}}_1)] &= 0 \\
\Leftrightarrow \frac{1}{u'(\underline{\underline{w}}_1)} &= \lambda_1.
\end{aligned} \tag{A.15}$$

From the previous four equations we obtain that

$$\begin{aligned}
\frac{1}{u'(\overline{\overline{w}}_i)} = \frac{1}{u'(\overline{w}_i)} = \frac{1}{u'(\underline{w}_i)} = \frac{1}{u'(\underline{\underline{w}}_i)} \\
\Leftrightarrow \overline{\overline{w}}_i = \overline{w}_i = \underline{w}_i = \underline{\underline{w}}_i = w_i.
\end{aligned} \tag{A.16}$$

Making use of (A.16), we take the derivative with respect to e_1:

$$\begin{aligned}
&\pi_j \cdot (\rho'(e_1) - \rho'(e_1) \cdot \rho(e_2)) + \\
\frac{\partial \mathcal{L}}{\partial e_1} = &\lambda_1 \cdot [-1] = 0 \\
\Leftrightarrow \rho'(e_1) \cdot \pi_j \cdot (1-\rho(e_2)) &= \lambda_1.
\end{aligned} \tag{A.17}$$

The same can be done for e_2. From equations (A.12) to (A.15) and (A.17) one can easily obtain

$$\rho'(e_1) \cdot \pi_j \cdot (1-\rho(e_2)) = \frac{1}{u'(w_1)} \tag{A.18}$$

and

$$w_i = u^{-1}(e_i). \tag{A.19}$$

Diversified efforts:

From the maximization problem we obtain the Lagrangian

$$\begin{aligned}
\mathcal{L} = \pi_m \cdot \pi_o \cdot [\rho(e_1) \cdot \rho(e_2) \cdot (1 - \overline{\overline{w}}_1 - \overline{\overline{w}}_2) + \\
\rho(e_1) \cdot (1 - \rho(e_2)) \cdot (1 - \overline{w}_1 - \underline{w}_2) + \\
(1 - \rho(e_1)) \cdot \rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + \\
(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
\pi_m \cdot (1 - \pi_o) \cdot [\rho(e_1) \cdot (1 - \overline{w}_1 - \underline{w}_2) + (1 - \rho(e_1)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
(1 - \pi_m) \cdot \pi_o \cdot [\rho(e_2) \cdot (1 - \underline{w}_1 - \overline{w}_2) + (1 - \rho(e_2)) \cdot (-\underline{\underline{w}}_1 - \underline{\underline{w}}_2)] + \\
(1 - \pi_m) \cdot (1 - \pi_o) \cdot [-\underline{\underline{w}}_1 - \underline{\underline{w}}_2] + \\
\lambda_1 \cdot [\pi_m \cdot \pi_o \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\
\rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\
(1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\
(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\
\pi_m \cdot (1 - \pi_o) \cdot [\rho(e_1) \cdot u(\overline{w}_1) + (1 - \rho(e_1)) \cdot u(\underline{\underline{w}}_1)] + \\
(1 - \pi_m) \cdot \pi_o \cdot [\rho(e_2) \cdot u(\underline{w}_1) + (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\
(1 - \pi_m) \cdot (1 - \pi_o) \cdot [u(\underline{\underline{w}}_1)] - e_1] + \\
\lambda_2 \cdot [\pi_m \cdot \pi_o \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \\
\rho(e_1) \cdot (1 - \rho(e_2)) \cdot u(\underline{w}_2) + \\
(1 - \rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + \\
(1 - \rho(e_1)) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
\pi_m \cdot (1 - \pi_o) \cdot [\rho(e_1) \cdot u(\underline{w}_2) + (1 - \rho(e_1)) \cdot u(\underline{\underline{w}}_2)] + \\
(1 - \pi_m) \cdot (\pi_o) \cdot [\rho(e_2) \cdot u(\overline{w}_2) + (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\
(1 - \pi_m) \cdot (1 - \pi_o) \cdot [u(\underline{\underline{w}}_2)] - e_2] = 0.
\end{aligned} \tag{A.20}$$

Taking derivatives with respect to the different wage levels for agent 1 yields

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \overline{\overline{w}}_1} = \lambda_1 \cdot [\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot (-1) + \\
\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] = 0 \\
\Leftrightarrow \frac{1}{u'(\overline{\overline{w}}_1)} = \lambda_1,
\end{aligned} \tag{A.21}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \overline{w}_1} = & (\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot (1-\rho(e_2)) + \pi_m \cdot (1-\pi_o) \cdot \rho(e_1)) \cdot (-1) + \\ & \lambda_1 \cdot [(\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot (1-\rho(e_2)) + \\ & \pi_m \cdot (1-\pi_o) \cdot \rho(e_1)) \cdot u'(\overline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\overline{w}_1)} = \lambda_1,\end{aligned} \tag{A.22}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{w}_1} = & (\pi_m \cdot \pi_o \cdot (1-\rho(e_1)) \cdot \rho(e_2) + (1-\pi_m) \cdot \pi_o \cdot \rho(e_2)) \cdot (-1) + \\ & \lambda_1 \cdot [(\pi_m \cdot \pi_o \cdot (1-\rho(e_1)) \cdot \rho(e_2) + \\ & + (1-\pi_m) \cdot \pi_o \cdot \rho(e_2)) \cdot u'(\underline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\underline{w}_1)} = \lambda_1,\end{aligned} \tag{A.23}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{\underline{w}}_1} = \lambda_1 \cdot [& (\pi_m \cdot \pi_o \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) + \\ & \pi_m \cdot (1-\pi_o) \cdot (1-\rho(e_1)) + \\ & (1-\pi_m) \cdot \pi_o \cdot (1-\rho(e_2)) + \\ & (1-\pi_m) \cdot (1-\pi_o)) \cdot (-1) + \\ & (\pi_m \cdot \pi_o \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) + \\ & \pi_m \cdot (1-\pi_o) \cdot (1-\rho(e_1)) + \\ & (1-\pi_m) \cdot \pi_o \cdot (1-\rho(e_2)) + \\ & (1-\pi_m) \cdot (1-\pi_o)) \cdot (u'(\underline{\underline{w}}_1))] = 0 \\ & \Leftrightarrow \frac{1}{u'(\underline{\underline{w}}_1)} = \lambda_1.\end{aligned} \tag{A.24}$$

The previous four equations are equivalent to (A.16). The optimal wage levels for agent 2 can be derived in a similar way.

Taking the derivative with respect to e_1 results in

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial e_1} = & \pi_m \cdot \rho'(e_1) \cdot (\pi_o \cdot (1-\rho(e_2)) + (1-\pi_o)) + \lambda_1 \cdot [-1] = 0 \\ & \Leftrightarrow \pi_m \cdot \rho'(e_1) \cdot (1 - \pi_o \cdot \rho(e_2)) = \lambda_1.\end{aligned} \tag{A.25}$$

From equations (A.21) to (A.24) and (A.25) one can easily obtain

$$\pi_m \cdot \rho'(e_1) \cdot (1 - \pi_o \cdot \rho(e_2)) = \frac{1}{u'(w_1)}. \tag{A.26}$$

The same can be done for e_2, which yields

$$\pi_o \cdot \rho'(e_2) \cdot (1 - \pi_m \cdot \rho(e_1)) = \frac{1}{u'(w_2)}. \tag{A.27}$$

Moreover, equation (A.16) also holds for the case of diversified efforts.

Asymmetric Information, n=2

Concentrated Efforts, Moral Hazard I:

The incentive compatibility constraint for agent 1 (likewise for agent 2) is given as

$$\begin{aligned} e_1 \in argmax_{\hat{e}_1} \pi_m \cdot [&\rho(\hat{e}_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ &\rho(\hat{e}_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\ &-\rho'(\hat{e}_1) \cdot \rho(e_2) \cdot u(\underline{w}_1) - \\ &\rho'(\hat{e}_1) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] - 1. \end{aligned} \tag{IC II: SB1 CE}$$

Therefore we add

$$\begin{aligned} \mu_i \cdot [\pi_m \cdot [&\rho'(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ &\rho'(e_1) \cdot (1 - \rho(e_2)) \cdot u(\overline{w}_1) + \\ &-\rho'(e_1) \cdot \rho(e_2) \cdot u(\underline{w}_1) - \\ &\rho'(e_1) \cdot (1 - \rho(e_2)) \cdot u(\underline{\underline{w}}_1)] - 1] \end{aligned} \tag{A.28}$$

to the left-hand side of the original Lagrange function (A.11) to obtain the updated Lagrangian. We take derivatives with respect to the different wage levels of agent 1 (likewise for agent 2):

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \overline{\overline{w}}_1} = &\pi_m \cdot \rho(e_1) \cdot \rho(e_2) \cdot (-1) + \\ &\lambda_1 \cdot [\pi_m \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] + \\ &\mu_1 \cdot [\pi_m \cdot \rho'(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] = 0 \\ &\Leftrightarrow \frac{1}{u'(\overline{\overline{w}}_1)} = \lambda_1 + \mu_1 \cdot \frac{\rho'(e_1)}{\rho(e_1)}, \end{aligned} \tag{A.29}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \overline{w}_1} = & \pi_m \cdot \rho(e_1) \cdot (1-\rho(e_2)) \cdot (-1) + \\ & \lambda_1 \cdot [\pi_m \cdot \rho(e_1) \cdot (1-\rho(e_2)) \cdot u'(\overline{w}_1)] + \\ & \mu_i \cdot [\pi_m \cdot \rho'(e_1) \cdot (1-\rho(e_2)) \cdot u'(\overline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\overline{w}_1)} = \lambda_1 + \mu_1 \cdot \frac{\rho'(e_1)}{\rho(e_1)},\end{aligned} \tag{A.30}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{w}_1} = & \pi_m \cdot (1-\rho(e_1)) \cdot \rho(e_2) \cdot (-1) + \\ & \lambda_1 \cdot [\pi_m \cdot (1-\rho(e_1)) \cdot \rho(e_2) \cdot u'(\underline{w}_1)] + \\ & \mu_1 \cdot [\pi_m \cdot (-\rho'(e_1)) \cdot \rho(e_2) \cdot u'(\underline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\underline{w}_1)} = \lambda_1 - \mu_1 \cdot \frac{\rho'(e_1)}{1-\rho(e_2)},\end{aligned} \tag{A.31}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{\underline{w}}_1} = & (\pi_m \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) + (1-\pi_m)) \cdot (-1) + \\ & \lambda_1 \cdot [(\pi_m \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) + (1-\pi_j)) \cdot u'(\underline{\underline{w}}_1)] + \\ & \mu_i \cdot [\pi_m \cdot (-\rho'(e_1)) \cdot (1-\rho(e_2)) \cdot u'(\underline{\underline{w}}_1)] \\ \Leftrightarrow \frac{1}{u'(\underline{\underline{w}}_1)} = & \lambda_1 - \mu_1 \cdot \frac{\pi_m \cdot \rho'(e_1) \cdot (1-\rho(e_2))}{\pi_m \cdot (1-\rho(e_1)) \cdot (1-\rho(e_2)) + (1-\pi_m)}.\end{aligned} \tag{A.32}$$

Diversified Efforts, Moral Hazard I:

The incentive compatibility constraint for agent 1 (likewise for agent 2) is given as

$$\begin{aligned} e_1 \in argmax_{\hat{e}_1} \pi_m \cdot \pi_o \cdot [\rho(\hat{e}_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ \rho(\hat{e}_1) \cdot (1-\rho(e_2)) \cdot u(\overline{w}_1) + \\ (1-\rho(\hat{e}_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\ (1-\rho(\hat{e}_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ (\pi_m \cdot (1-\pi_o) \cdot [\rho(\hat{e}_1) \cdot u(\overline{w}_1) + (1-\rho(\hat{e}_1)) \cdot u(\underline{\underline{w}}_1)] + \\ (1-\pi_m) \cdot \pi_o \cdot [\rho(e_2) \cdot u(\underline{w}_1) + (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ (1-\pi_m) \cdot (1-\pi_o) \cdot [u(\underline{\underline{w}}_1)] - \hat{e}_1. \end{aligned} \tag{IC II: SB1 DE}$$

We add

$$\begin{aligned} \mu_i \cdot [\pi_m \cdot \pi_o \cdot [\rho'(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_1) + \\ \rho(e_1) \cdot (1-\rho(e_2)) \cdot u(\overline{w}_1) + \\ (1-\rho(\hat{e}_1)) \cdot \rho(e_2) \cdot u(\underline{w}_1) + \\ (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ (\pi_m \cdot (1-\pi_o) \cdot [\rho(e_1) \cdot u(\overline{w}_1) + (1-\rho(e_1)) \cdot u(\underline{\underline{w}}_1)] + \\ (1-\pi_m) \cdot \pi_o \cdot [\rho(e_2) \cdot u(\underline{w}_1) + (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_1)] + \\ (1-\pi_m) \cdot (1-\pi_o) \cdot [u(\underline{\underline{w}}_1)] - 1] \end{aligned} \tag{A.33}$$

to the left-hand side of equation (A.20) to obtain an updated Lagrangian. We once more take derivatives with respect to the different wage levels of agent 1 (likewise for agent 2):

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \overline{\overline{w}}_1} = & \pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot (-1) + \\ & \lambda_1 \cdot [\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] + \\ & \mu_1 \cdot [\pi_m \cdot \pi_o \cdot \rho'(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\overline{\overline{w}}_1)} = \lambda_1 + \mu_1 \cdot \frac{\rho'(e_1)}{\rho(e_1}, \end{aligned} \tag{A.34}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \overline{w}_1} = & \pi_m \cdot \rho(e_1) \cdot (\pi_o \cdot (1-\rho(e_2)) + (1-\pi_o)) \cdot (-1) + \\ & \lambda_1 \cdot [\pi_m \cdot \rho(e_1) \cdot (\pi_o \cdot (1-\rho(e_2)) + (1-\pi_o)) \cdot u'(\overline{w}_1)] + \\ & \mu_1 \cdot [\pi_m \cdot \rho'(e_1) \cdot (\pi_o \cdot (1-\rho(e_2)) + (1-\pi_o)) \cdot u'(\overline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\overline{w}_1)} = \lambda_1 + \mu_1 \cdot \frac{\rho'(e_1)}{\rho(e_1)}, \end{aligned} \tag{A.35}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{w}_1} = & (\pi_m \cdot (1-\rho(e_1)) + (1-\pi_m)) \cdot \pi_o \cdot \rho(e_2) \cdot (-1) + \\ & \lambda_1 \cdot [(\pi_m \cdot (1-\rho(e_1)) + (1-\pi_m)) \cdot \pi_o \cdot \rho(e_2) \cdot u'(\underline{w}_1)] + \\ & \mu_1 \cdot [\pi_m \cdot \pi_o \cdot (-\rho'(e_1)) \cdot \rho(e_2) \cdot u'(\underline{w}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\underline{w}_1)} = \lambda_1 - \mu_1 \cdot \frac{\pi_m \cdot \rho'(e_1)}{1-\pi_m \cdot \rho(e_1)},\end{aligned} \tag{A.36}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \underline{\underline{w}}_1} = & (1-\pi_m \cdot \rho(e_1)) \cdot (1-\pi_o \cdot \rho(e_2)) \cdot (-1) + \\ & \lambda_1 \cdot [(1-\pi_m \cdot \rho(e_1)) \cdot (1-\pi_o \cdot \rho(e_2)) \cdot u'(\underline{\underline{w}}_1)] + \\ & \mu_1 \cdot [\pi_m \cdot (-\rho'(e_1)) \cdot (1-\pi_o \cdot \rho(e_2)) \cdot u'(\underline{\underline{w}}_1)] = 0 \\ & \Leftrightarrow \frac{1}{u'(\underline{\underline{w}}_1)} = \lambda_1 - \mu_1 \cdot \frac{\pi_m \cdot \rho'(e_1)}{1-\pi_m \cdot \rho(e_1)}.\end{aligned} \tag{A.37}$$

Diversified efforts, Moral Hazard II:

Condition (IC2 II: SB2 DE) must hold with equality (otherwise the principal would give away utility for free). We incorporate this constraint into the Lagrange function by adding

$$\begin{aligned}\nu_2 \cdot [& (\pi_m \cdot \pi_o) \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \\ & \rho(e_1) \cdot (1-\rho(e_2)) \cdot u(\underline{w}_2) + \\ & (1-\rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + \\ & (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\ & (\pi_m \cdot (1-\pi_o)) \cdot [\rho(e_1) \cdot u(\underline{w}_2) + (1-\rho(e_1)) \cdot u(\underline{\underline{w}}_2)] + \\ & ((1-\pi_m) \cdot \pi_o) \cdot [\rho(e_2) \cdot u(\overline{w}_2) + (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\ & ((1-\pi_m) \cdot (1-\pi_o)) \cdot u(\underline{\underline{w}}_2) - \\ & (\pi_m \cdot [\rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \\ & \rho(e_1) \cdot (1-\rho(e_2)) \cdot u(\underline{w}_2) + \\ & (1-\rho(e_1)) \cdot \rho(e_2) \cdot u(\overline{w}_2) + \\ & (1-\rho(e_1)) \cdot (1-\rho(e_2)) \cdot u(\underline{\underline{w}}_2)] + \\ & (1-\pi_m) \cdot u(\underline{\underline{w}}_2))]\end{aligned} \tag{A.38}$$

to the left-hand side of the former Lagrange function (equations (A.20) and (A.33)) and take derivatives with respect to the different wage levels of agent 2.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \overline{\overline{w}}_2} = & \quad \pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot (-1) + \\
& \lambda_2 \cdot [\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_2)] + \\
& \mu_2 \cdot [\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho'(e_2) \cdot u'(\overline{\overline{w}}_2)] + \\
& \nu_2 \cdot [\pi_m \cdot (\pi_o - 1) \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\overline{\overline{w}}_2)] = 0 \\
\Leftrightarrow & \frac{1}{u'(\overline{\overline{w}}_2)} = \lambda_2 + \mu_2 \cdot \frac{\rho'(e_2)}{\rho(e_2)} + \nu_2 \cdot \frac{\pi_o - 1}{\pi_o},
\end{aligned} \tag{A.39}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \overline{w}_2} = & \quad \pi_o \cdot \rho(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1)) + (1 - \pi_m)) \cdot (-1) + \\
& \lambda_2 \cdot [\pi_o \cdot \rho(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1)) + (1 - \pi_m)) \cdot u'(\overline{w}_2)] + \\
& \mu_1 \cdot [\pi_o \cdot \rho'(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1)) + (1 - \pi_m)) \cdot u'(\overline{w}_2)] + \\
& \nu_2 \cdot [\rho(e_2) \cdot (\pi_m \cdot (\rho(e_1) \cdot (1 - \pi_o) - 1) + \pi_o) \cdot u'(\overline{w}_2)] = 0 \\
\Leftrightarrow & \frac{1}{u'(\overline{w}_2)} = \lambda_2 + \mu_2 \cdot \frac{\rho'(e_2)}{\rho(e_2)} + \nu_2 \cdot \frac{\pi_m \cdot (\rho(e_1) \cdot (1 - \pi_o) - 1) + \pi_o}{\pi_o \cdot (1 - \pi_m \cdot \rho(e_1))},
\end{aligned} \tag{A.40}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \underline{w}_2} = & \quad (\pi_o \cdot (1 - \rho(e_2)) + (1 - \pi_o)) \cdot \pi_m \cdot \rho(e_1) \cdot (-1) + \\
& \lambda_2 \cdot [(\pi_o \cdot (1 - \rho(e_2)) + (1 - \pi_o)) \cdot \pi_m \cdot \rho(e_1) \cdot u'(\underline{w}_2)] + \\
& \mu_1 \cdot [\pi_o \cdot \pi_m \cdot \rho(e_1) \cdot (-\rho'(e_2)) \cdot u'(\underline{w}_2)] + \\
& \nu_2 \cdot [(1 - \pi_o) \cdot \pi_m \cdot \rho(e_1) \cdot \rho(e_2) \cdot u'(\underline{w}_2)] = 0 \\
\Leftrightarrow & \frac{1}{u'(\underline{w}_2)} = \lambda_1 - \mu_2 \cdot \frac{\pi_o \cdot \rho'(e_2)}{1 - \pi_o \cdot \rho(e_2)} + \nu_2 \cdot \frac{\rho(e_2) \cdot (1 - \pi_o)}{1 - \pi_o \cdot \rho(e_2)},
\end{aligned} \tag{A.41}$$

and

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \underline{\underline{w}}_2} = & \quad (1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2)) \cdot (-1) + \\
& \lambda_2 \cdot [(1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2)) \cdot u'(\underline{\underline{w}}_2)] + \\
& \mu_2 \cdot [\pi_m \cdot (-\rho'(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2)) \cdot u'(\underline{\underline{w}}_2)] + \\
& \nu_2 \cdot [\rho(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1)) \cdot (1 + \pi_o)) - \pi_o)] = 0 \\
\Leftrightarrow & \frac{1}{u'(\underline{\underline{w}}_2)} = \lambda_2 - \mu_2 \cdot \frac{\pi_m \cdot \rho'(e_1)}{1 - \pi_m \cdot \rho(e_1)} + \\
& \nu_2 \cdot \frac{\rho(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1) \cdot (1 - \pi_o)) - \pi_o)}{(1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2))}.
\end{aligned} \tag{A.42}$$

For agent 1, no additional constraint is necessary, since choosing the mainstream technology is in his personal interest. Hence, his respective wage levels are still determined by

conditions (A.34) to (A.37).

A.2 Proofs

Proof of Proposition 1.3

From the first-order conditions of the optimization problem (cf. Appendix A.1), we can conclude that the optimal effort and wage levels are implicitly defined by the following system of equations:

$$\rho'(e_1) \cdot \pi_j \cdot (1 - \rho(e_2)) = \frac{1}{u'(u^{-1}(e_1))} \tag{A.43}$$

and

$$\rho'(e_2) \cdot \pi_j \cdot (1 - \rho(e_1)) = \frac{1}{u'(u^{-1}(e_2))}. \tag{A.44}$$

Due to our technical assumptions, especially condition (1.13), it is guaranteed that an interior solution will exist for equation (A.43) as long as $e_2 < e^*$. Likewise, an interior solution for equation (A.44) will exist whenever $e_1 < e^*$. For $0 < e_2 < e^*$, equation (A.43) implies levels of e_1 that are smaller than e^*. Similarly, for $0 < e_1 < e^*$, equation (C.15) implies levels of e_2 that are smaller than e^*. Therefore, we can conclude that $0 < e_1, e_2 < e^*$.

Equations (A.43) and (A.44) can be rearranged to

$$\rho(e_2) = 1 - \frac{1}{u'(u^{-1}(e_1)) \cdot \rho'(e_1) \cdot \pi_j} \tag{A.45}$$

and

$$\rho(e_1) = 1 - \frac{1}{u'(u^{-1}(e_2)) \cdot \rho'(e_2) \cdot \pi_j}. \tag{A.46}$$

Since $\rho(\cdot)$ is an invertible function, we can substitute the inverse function of $\rho(e_1)$ into (A.44) and obtain

$$\begin{aligned} \rho' \left(\rho^{-1} \left(1 - \frac{1}{u' \left(u^{-1} \left(e_1 \right) \right) \cdot \rho\,'(e_1) \cdot \pi_j} \right) \right) \cdot \pi_j \cdot \left(1 - \rho \left(e_1 \right) \right) \\ = \frac{1}{u' \left(u^{-1} \left(\rho^{-1} \left(1 - \frac{1}{u'(u^{-1}(e_1)) \cdot \rho'(e_1) \cdot \pi_j} \right) \right) \right)}. \end{aligned} \tag{A.47}$$

For $e_1 = 0$, the left-hand side of equation (A.47) is smaller than the right-hand side whenever $1 - \frac{1}{u'(u^{-1}(0)) \cdot \rho'(0) \cdot \pi_j} > \rho(e^*)$. This inequality can be rearranged to condition

(1.13) and is true by assumption. For $e_1 = e^*$, the left-hand side can be rewritten as $\rho'(0) \cdot \pi_j \cdot (1 - \rho(e^*))$, and the right-hand side becomes $\frac{1}{u'(0)}$. Clearly, again by assumption (1.13), the left-hand side must be larger than the right-hand side. Since the left-hand side is strictly increasing and the right-hand side is strictly decreasing in e_1, a unique solution for equation (A.47) must exist.

Likewise, we substitute the inverse function of $\rho(e_2)$ into (A.43) and obtain

$$\begin{aligned} \rho' \left(\rho^{-1} \left(1 - \frac{1}{u' \left(u^{-1} \left(e_2 \right) \right) \cdot \rho\,'(e_2) \cdot \pi_j} \right) \right) \cdot \pi_j \cdot \left(1 - \rho \left(e_2 \right) \right) \\ = \frac{1}{u' \left(u^{-1} \left(\rho^{-1} \left(1 - \frac{1}{u'(u^{-1}(e_2)) \cdot \rho'(e_2) \cdot \pi_j} \right) \right) \right)}. \end{aligned} \tag{A.48}$$

A direct implication of equations (A.47) and (A.48) is that $e_1 = e_2$.

What still needs to be shown is that the second-order conditions for an optimum are satisfied. To avoid the construction of a Hessian matrix which would render the problem more intricate, we make use of the fact that $e_1 = e_2$ and transform the multivariate optimization problem into a univariate one. We also exploit the fact that $\underline{\underline{w}}_i = \underline{w}_i = \overline{w}_i = \overline{\overline{w}}_i = w_i$. Then, letting $w(e_i)$ denote the (expected) wage level as a function of e_i, the optimization problem essentially becomes

$$\pi_j \cdot (\rho(e_i) + (1 - \rho(e_i)) \cdot \rho(e_i)) - 2 \cdot w(e_i) \tag{A.49}$$

which is clearly a concave problem, as the second-order condition is

$$\pi_j \cdot (2 \cdot \rho''(e_i) \cdot (1 - \rho(e_i)) - \rho'(e_i) \cdot 2 \cdot \rho'(e_i)) - 2 \cdot w''(e_i) < 0. \tag{A.50}$$

Since symmetric effort is optimal, we can more easily derive the optimal effort level by solving equation (1.17). □

Proof of Proposition 1.4

Condition (1.20) holds if and only if

$$\begin{aligned}(\pi_m \cdot \pi_o) \cdot (\rho(e_1') \cdot \rho(e_2') + \rho(e_1') \cdot (1 - \rho(e_2')) + (1 - \rho(e_1')) \cdot \rho(e_2')) + \\ \pi_m \cdot (1 - \pi_o) \cdot \rho(e_1') + (1 - \pi_m) \cdot \pi_o \cdot \rho(e_2') - w_1' - w_2' > \\ \pi_m \cdot (2 \cdot \rho(e_1) - \rho(e_1)^2) - 2 \cdot w_1 \\ \Leftrightarrow \pi_o > \frac{\pi_m \cdot (\rho(e_1) \cdot (2 - \rho(e_1)) - \rho(e_1')) - 2 \cdot w_1 + w_1' + w_2'}{\rho(e_2') \cdot (1 - \pi_m \cdot \rho(e_1'))},\end{aligned} \tag{A.51}$$

where we use the fact that $e_1 = e_2$ and $w_1 = w_2$.

Next, we show that for $0 > \pi_m > 1$, $\tilde{\pi}_o$ is strictly larger than 0. We do so by first showing that for $\pi_o = 0$, concentrated efforts are strictly better than diversified efforts. Since $\rho(e_2') = 0$ for $\pi_o = 0$, the expected payoff for diversified efforts equals the expected payoff of a single researcher, using the mainstream technology. From condition (1.13) we know that every expected return of a single researcher can be obtained more cheaply with two researchers both using the same technology. Hence, $\tilde{\pi}_o$ must be larger than zero.

Second, we show that for $\pi_o = \pi_m$, diversified efforts are strictly better, such that $\tilde{\pi}_o$ is strictly smaller than π_m. We plug the optimal effort-wage combination for concentrated efforts into $E(V_{mo}(\cdot))$ and yield

$$\begin{aligned}1 - (1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_m \cdot \rho(e_1)) - 2 \cdot w_1 > \\ \pi_m \cdot (1 - (1 - \rho(e_1)) \cdot (1 - \rho(e_1))) - 2 \cdot w_1 \\ \Leftrightarrow \pi_m \cdot \rho(e_1) \cdot (2 - \pi_m \cdot \rho(e_1)) > \pi_m \cdot \rho(e_1) \cdot (2 - \rho(e_1)) \\ \Leftrightarrow 1 > \pi_m.\end{aligned} \tag{A.52}$$

Again, this condition is always satisfied, such that $\tilde{\pi}_o < \pi_m$.

Lastly, we show that $E(V_{mo}(\cdot))$ is strictly increasing in π_o and $E(V_{mm}(\cdot))$ is not affected by changes of π_o, which implies that a unique intersection of both payoff functions must exist. If π_o increases, but the effort-wage combination remains unchanged, $E(V_{mo}(\cdot))$ rises. Hence, increasing the effort when π_o rises must necessarily yield weakly higher returns than keeping the effort level constant, such that $E(V_{mo}(\cdot))$ is strictly increasing in π_o. According to equation (1.17), $E(V_{mm}(\cdot))$ does not depend on π_o, such that the intersection must be unique. □

Proof of Proposition 1.5

The existence of a unique threshold $\tilde{\pi}_o^{SB1}$ can be proven using an argument much like the one used to prove Proposition 1.4 . What remains to be shown is that $\tilde{\pi}_o^{SB1} > \tilde{\pi}_o$.

To do so, we first define an upper bound for $\tilde{\pi}_o$, which we will subsequently compare to a lower bound for $\tilde{\pi}_o^{SB1}$. First, however, we will establish some useful results on the relative size of different wage and effort levels. Due to the agents' risk aversion, it must be the case that

$$w'(e_i) < w_m^{SB1\prime}(e_i) \leq w_o^{SB1\prime}(e_i) \tag{A.53}$$

for all e_i. Here, $w'(\cdot)$ and $w_j^{SB1\prime}(\cdot)$, respectively, denote the *marginal* wage levels for symmetric and asymmetric information, when using technology j. A given level of effort is cheaper to induce under symmetric information than under asymmetric information with technology m. Under asymmetric information, any effort level is weakly cheaper to induce with technology m than with technology o, and the inequality is strict whenever $\pi_m > \pi_o$.

Condition (A.53) directly implies that in optimum

$$e_i > e_i^{SB1} \tag{A.54}$$

must hold true. It remains unclear, however, whether $e_i^{'} > e_i^{'SB1}$ is also satisfied for all i.[8] In the remainder of the proof, we therefore analyze the cases $e_1^{'} > e_1^{'SB1}$ and $e_1^{'} \leq e_1^{'SB1}$ separately.

We start by analyzing $e_i^{'} > e_i^{'SB1}$. From the principal's perspective, the marginal gain from letting agent 2 work with the mainstream technology rather than the outsider technology is weakly higher if

$$\pi_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'})\right) \leq \pi_m \cdot \rho'(e_2) \cdot (1 - \rho(e_1)) \,. \tag{A.55}$$

Note that in the above inequality, $e_1^{'}$ and e_1 denote the first-best optimal levels of the

[8] To see this, recall that a *given* effort level of agent 2 implies a lower optimal effort level of agent 1 under asymmetric information. However, the effort level of agent 2 will also adjust under asymmetric information (and possibly shrink as compared to symmetric information), which causes $e_1^{'SB1}$ to increase. Since we have two opposing effects, without further specifications it remains unresolved which effect will eventually prevail.

first agent's effort level for diversified efforts and concentrated efforts, respectively, under the assumption that the second agent's effort levels are also optimally chosen. Keeping in mind our assumption that $e_1^{'} > e_1^{'SB1}$, the inequality can be replaced by the more restrictive version

$$\begin{aligned} \tilde{\pi}_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'SB1})\right) &\leq \pi_m \cdot \rho'(e_2) \cdot (1 - \rho(e_1)) \\ \Leftrightarrow \tilde{\pi}_o &\leq \frac{\pi_m \cdot (1 - \rho(e_1))}{1 - \pi_m \cdot \rho(e_1^{'SB1})}. \end{aligned} \tag{A.56}$$

Likewise, under asymmetric information, the marginal gain from letting agent 2 work with the outsider technology rather than the mainstream technology is weakly higher if

$$\pi_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'SB1})\right) > \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB1})\right). \tag{A.57}$$

We replace the above inequality with the more restrictive condition

$$\begin{aligned} \tilde{\pi}_o^{SB1} \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'SB1})\right) &> \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB1})\right) \\ \Leftrightarrow \tilde{\pi}_o^{SB1} &> \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{'SB1})}. \end{aligned} \tag{A.58}$$

$\tilde{\pi}_o^{SB1} > \tilde{\pi}_o$ must be true whenever the following condition is satisfied:

$$\begin{aligned} \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{'SB1})} &> \frac{\pi_m \cdot (1 - \rho(e_1))}{1 - \pi_m \cdot \rho(e_1^{'SB1})} \\ \Leftrightarrow \rho(e_1) &> \rho(e_1^{SB}). \end{aligned} \tag{A.59}$$

Inequality (A.59) is satisfied because of (A.54).

We proceed by analyzing $e_i^{'} \leq e_i^{'SB}$. Condition (A.55) can be replaced by

$$\begin{aligned} \tilde{\pi}_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'})\right) &\leq \pi_m \cdot \rho'(e_2) \cdot (1 - \rho(e_1)) \\ \Leftrightarrow \tilde{\pi}_o &\leq \frac{\pi_m \cdot (1 - \rho(e_1))}{1 - \pi_m \cdot \rho(e_1^{'})}. \end{aligned} \tag{A.60}$$

Moreover, condition (A.57) can be substituted by

$$\begin{aligned} \tilde{\pi}_o^{SB1} \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{'})\right) &> \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB1})\right) \\ \Leftrightarrow \tilde{\pi}_o^{SB1} &> \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{'})}. \end{aligned} \tag{A.61}$$

Once more, $\tilde{\pi}_o^{SB1} > \tilde{\pi}_o$ is true whenever the following condition is satisfied:

$$\begin{aligned} \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{'})} &> \frac{\pi_m \cdot (1 - \rho(e_1))}{1 - \pi_m \cdot \rho(e_1^{'})} \\ &\Leftrightarrow \rho(e_i) > \rho(e_i^{SB1}). \end{aligned} \tag{A.62}$$

Since (A.62) is always satisfied, the proof is completed. □

Proof of Proposition 1.6

Since only two wage levels have to be considered (Lemma 1.2), agent i prefers to choose the mainstream technology if

$$\begin{aligned} \pi_m \cdot \rho(e_i') \cdot u(\overline{w}_i') + (1 - \pi_m \cdot \rho(e_i')) \cdot u(\underline{w}_i') \geq \\ \pi_o \cdot \rho(e_i') \cdot u(\overline{w}_i)' + (1 - \pi_o \cdot \rho(e_i')) \cdot u(\underline{w}_i') \\ \Leftrightarrow \pi_m \geq \pi_o. \end{aligned} \tag{A.63}$$

Hence agent 2 will always deviate. □

Proof of Lemma 1.3

For $\overline{w}_2^{SB2} > \overline{\overline{w}}_2^{SB2}$ to be true, equations (A.34) and (A.40) imply that it is sufficient to show that

$$\begin{aligned} \frac{\pi_m \cdot (\rho(e_1) \cdot (1 - \pi_o) - 1) + \pi_o}{\pi_o \cdot (1 - \pi_m \cdot \rho(e_1))} &> \frac{\pi_o - 1}{\pi_o} \\ \Leftrightarrow 0 > (\pi_o - 1) \cdot (1 - \pi_m \cdot \rho(e_1)) - (\pi_m \cdot (\rho(e_1) \cdot (1 - \pi_o) - 1) + \pi_o) \\ \Leftrightarrow 1 > \pi_m. \end{aligned} \tag{A.64}$$

Likewise, for $\underline{w}_2^{SB2} > \underline{\overline{w}}_2^{SB2}$ to hold true, equations (A.41) and (A.42) imply

$$\begin{aligned} \frac{\rho(e_2) \cdot (1 - \pi_o)}{1 - \pi_o \cdot \rho(e_2)} &> \frac{\rho(e_2) \cdot (\pi_m \cdot (1 - \rho(e_1) \cdot (1 - \pi_o)) - \pi_o)}{(1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2))} \\ \Leftrightarrow (1 - \pi_o) \cdot (1 - \pi_m \cdot \rho(e_1)) - (\pi_m \cdot (1 - \rho(e_1) \cdot (1 - \pi_o)) - \pi_o) > 0 \\ \Leftrightarrow 1 > \pi_m. \quad \square \end{aligned} \tag{A.65}$$

Proof of Proposition 1.7

Since $u(\cdot)$ is concave, it is true that

$$\begin{aligned}
&\pi_m \cdot \pi_o \cdot \rho(e_1) \cdot \rho(e_2) \cdot u(\overline{\overline{w}}_2) + \\
&\pi_m \cdot \rho(e_1) \cdot (1 - \pi_o \cdot \rho(e_2)) \cdot u(\underline{w}_2) + \\
&(1 - \pi_m \cdot \rho(e_1)) \cdot \pi_o \cdot \rho(e_2)) \cdot u(\overline{w}_2) + \\
&(1 - \pi_m \cdot \rho(e_1)) \cdot (1 - \pi_o \cdot \rho(e_2)) \cdot u(\underline{\underline{w}}_2) = e_2 < \\
&\pi_o \cdot \rho(e_2) \cdot u(\pi_m \cdot \rho(e_1) \cdot \overline{\overline{w}}_2 + (1 - \pi_m \cdot \rho(e_1)) \cdot \overline{w}_2) + \\
&(1 - \pi_o \cdot \rho(e_2)) \cdot u(\pi_m \cdot \rho(e_1) \cdot \underline{w}_2 + (1 - \pi_m \cdot \rho(e_1)) \cdot \underline{\underline{w}}_2).
\end{aligned} \tag{A.66}$$

If $(\overline{\overline{w}}_2^{SB2}, \overline{w}_2^{SB2}, \underline{w}_2^{SB2}, \underline{\underline{w}}_2^{SB2})$ are the solutions to the principal's optimization problem under Moral Hazard II, the left-hand side of (A.66) equals e_2, as agent 2's participation constraint is binding. Under Moral Hazard I, the principal conditions agent 2's wage only on his own success. Keeping the expected value fixed, the principal can adjust the spread between payments so that the agent is incentivized to provide the same effort. Thus, she can achieve the same likelihood of success at a lower cost. □

Proof of Proposition 1.8
Let e_i^{SB2} and $E(W_i^{SB2})$ denote the optimal effort and expected wage levels for concentrated efforts, and let $e_i^{'SB2}$ and $E(w_i^{'SB2})$ denote the optimal effort and expected wage levels for diversified efforts when the effort level and technology choice are unobservable. Then, a revised form of condition (A.51) yields

$$\pi_o > \frac{\pi_m \cdot (\rho(e_1^{SB2}) \cdot (2 - \rho(e_1^{SB2})) - \rho(e_1^{'SB2})) - 2 \cdot E(W_1^{SB2}) + E(W_1^{'SB2}) + E(W_2^{'SB2})}{\rho(e_2^{'SB2}) \cdot (1 - \pi_m \cdot \rho(e_1^{'SB2}))}. \tag{A.67}$$

Such an intersection is guaranteed to exist whenever $\tilde{\pi}_o^{SB2} < \pi_m$. To show that this is true, we assume $\pi_o = \pi_m$ and plug the optimal effort-wage combination for $E(V_{mm}^{SB2}(\cdot))$ into $E(V_{mo}^{SB2}(\cdot))$ and compare payoffs. We yield a revised form of inequality (A.52):

$$\begin{aligned}
&1 - (1 - \pi_m \cdot \rho(e_1^{SB2})) \cdot (1 - \pi_m \cdot \rho(e_1^{SB2})) - 2 \cdot E(W_1^{'SB2}) > \\
&\pi_m \cdot (1 - (1 - \rho(e_1^{SB2})) \cdot (1 - \rho(e_1^{SB2}))) - 2 \cdot E(W_1^{'SB2}) \\
\Leftrightarrow\ &\pi_m \cdot \rho(e_1^{SB2}) \cdot (2 - \pi_m \cdot \rho(e_1^{SB2})) > \pi_m \cdot \rho(e_1^{SB2}) \cdot (2 - \rho(e_1^{SB2})) \\
&\Leftrightarrow 1 > \pi_m.
\end{aligned} \tag{A.68}$$

Proof that $\tilde{\pi}_o^{SB2} > \tilde{\pi}_o^{SB1}$ is structured in mich the same way as proof of Proposition 1.5. First, it is worth recalling that that proof of Proposition 1.7 implies that

$$w_m^{SB1\prime}(e_1) = w_m^{SB2\prime}(e_1), \quad w_o^{SB1\prime}(e_2) < w_o^{SB2\prime}(e_2) \tag{A.69}$$

where $w_j^{SBX\prime}(e_i)$ describes the marginal expected wage of agent i, employing technology j. Thisdirectly implies that in optimum

$$e_1^{\prime SB2} > e_1^{\prime SB1}, \quad e_2^{\prime SB2} < e_2^{\prime SB1}. \tag{A.70}$$

Under Moral Hazard I, from the principal's perspective the marginal gain from letting agent 2 work with technology m (instead of technology o) is weakly higher if

$$\pi_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{\prime SB1})\right) \leq \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB1})\right). \tag{A.71}$$

This inequality is replaced by the more restrictive version

$$\begin{aligned} \tilde{\pi}_o^{SB1} \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{\prime SB1})\right) &\leq \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB1})\right) \\ \Leftrightarrow \tilde{\pi}_o^{SB1} &\leq \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{\prime SB1})}. \end{aligned} \tag{A.72}$$

Likewise, under Moral Hazard II, the principal will let the second agent work with technology o (instead of technology m) if and only if

$$\pi_o \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{\prime SB2})\right) > \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB2})\right). \tag{A.73}$$

Once more, we replace the original inequality with a more restrictive version:

$$\begin{aligned} \tilde{\pi}_o^{SB2} \cdot \rho'(e_2) \cdot \left(1 - \pi_m \cdot \rho(e_1^{\prime SB2})\right) &> \pi_m \cdot \rho'(e_2) \cdot \left(1 - \rho(e_1^{SB2})\right) \\ \Leftrightarrow \tilde{\pi}_o^{SB2} &> \frac{\pi_m \cdot (1 - \rho(e_1^{SB2}))}{1 - \pi_m \cdot \rho(e_1^{\prime SB2})}. \end{aligned} \tag{A.74}$$

For $\tilde{\pi}_o^{SB2} > \tilde{\pi}_o^{SB1}$ to be true, the following inequality must necessarily be satisfied:

$$\frac{\pi_m \cdot (1 - \rho(e_1^{SB2}))}{1 - \pi_m \cdot \rho(e_1^{\prime SB2})} > \frac{\pi_m \cdot (1 - \rho(e_1^{SB1}))}{1 - \pi_m \cdot \rho(e_1^{\prime SB1})} \Leftrightarrow \rho(e_1^{\prime SB2}) > \rho(e_1^{\prime SB1}) \tag{A.75}$$

Since $e_1^{SB1} = e_1^{SB2}$ and because of condition (A.70), the inequality must be satisfied. □

2 Strategic Delay in R&D Projects - An Agency Perspective*

Elisabeth Schulte
University of Marburg

Matthias Verbeck†
University of Marburg

Abstract

We show that strategic delay can pose a problem in delegated R&D projects. In our model, a principal delegates a research project to an agent. Depending on the agent's effort provision in two time periods, the research project can be completed either early, late or never. Our central assumption is that the agent is able to opportunistically withhold possible early completion from the principal (strategic delay). We derive the conditions under which strategic delay poses a problem. There are two options for the contract's optimal adjustment that both fall short of the first-best solution. (1) The contract prevents strategic delay by separating between successful and unsuccessful agents after period 1, but thereby distorts the agent's working incentives in both periods. (2) The principal strategically delays the start of the research project until the second period. We discuss several model extensions and possible institutional remedies to mitigate the problem.

Keywords: Principal-agent problem, Moral hazard, Adverse selection, Information disclosure, Project completion, Strategic delay, Incentives in research and development

JEL codes: D82, D83, D86, M52

*Part of the research was conducted at University of Texas at Dallas. Their hospitality is gratefully acknowledged.

†Corresponding author: matthias.verbeck@gmail.com

2.1 Introduction

The standard moral hazard model in the fashion of Holmström (1979) has contributed enormously to understanding efficiency losses in agency relationships, i.e. when tasks are delegated rather than performed by the bearer of the payoff consequences. One of the pivotal premises of the model is that the agent's actions are hidden, whereas the output, which is assumed to be stochastically related to the agent's actions, is common knowledge. The notion that the observability of output is unproblematic seems especially uncontroversial in the case of a one-shot interaction between the two parties. Even if an output is technically hidden to the principal (an assumption that is valid for many agency relationships), the reported output can still be verified by the principal. An agent in an optimal one-shot contract will see his wage increase with performance. Hence, the agent would never underreport his level of output, thus rationalizing the common-knowledge assumption.

However, when the interaction between principal and agent lasts more than one period, this simple reasoning becomes more questionable. We consider a two-period timescale in which an agent is supposed to work on a project and the agent's success (i.e. project completion) can occur early (period 1), late (period 2) or never. One characteristic of this scenario is that the desired effort level in the second period critically hinges on the first period's outcome, because only the case of a first period failure would arouse the principal's interest in a positive effort level for the second period. A rational agent who seeks to maximize his income might therefore be reluctant to immediately disclose an early success, anticipating that his or her second period effort (possibly resulting in an extra payment) will only be required in the case of a (reported) failure.

The unobservability of the timing of success may give rise to a conflict between creating working incentives on the one hand, and ensuring truthful reporting on the other. Where it is advantageous from the agent's perspective, s/he may then *strategically delay* the project's completion by leading the principal to believe that a project already finalized in period 1 was not actually completed before the end of period 2.

Indeed, projects finalized later than originally planned are the rule rather than the exception. Examples of projects that were eventually completed with delays of up to several years exist in abundance. According to Parkinson's law, "work expands so as to fill the

time available for its completion"(Parkinson (1957)), an observation that will be familiar to anybody with experience in time-critical undertakings. Missed deadlines and unforeseen delays are a notorious problem that plagues project managers, home-builders and doctoral advisors alike. Besides being a permanent cause of nuisance, they are also of high economic relevance.

Frequently cited explanations for this phenomenon stem from concepts rooted in behavioral economics (e.g. O'Donoghue and Rabin (1999)), such as limited self-control, time-inconsistent preferences and systematic overestimation of one's own abilities. Other rationales come from managerial economics, e.g. restrictive deadlines as a means of creating working incentives (e.g. Green and Taylor (2016)) or the strategic slowdown of projects due to the manager's intrinsic utility derived from holding the position of project leader (Katolnik and Schöndube (2019)). This paper aims at adding a novel perspective on the problem. The question we pose is whether the agent's informational advantage about a project's true completion status might be a (further) explanation for the omnipresence of belated project completions and violated deadlines in delegated research projects.

Our analysis intends to focus on *R&D projects*, which typically feature several characteristics that make them relatively unique and therefore worth investigating in their own right. We present a parsimonious economic model that captures these characteristics, and analyze the problem of delegated research from the perspective of principal-agent theory. For our analysis, we follow the model of Holmström (1979) with an agent who chooses a continuous level of unobservable effort and generates a binary output, stochastically linked to the agent's input. These assumptions are a good approximation of conditions found in many research settings, where a high degree of effort and dedication is a necessary, but not a sufficient condition for the success of the project, which is typically characterized by the ever-present possibility of failure. Thus, in research undertakings, effort is typically less deterministically related to the output, compared to other kinds of projects (e.g. building projects). Therefore, it seems natural to model the project's outcome as binary (success or failure), e.g. researchers in the pharmaceutical industry either succeed in making a new drug market-ready or not. What is more, if a research endeavor remained fruitless in a given period, the effort invested in that period has no more than a very limited influence, if any, on the probability of completing the undertaking at a later stage. Often,

research starts from scratch after a particular approach turns out to be a dead end. This characteristic distinguishes research projects from other kinds of projects where effort accumulates over time and early investments of effort do have an impact on the prospects of completing the project at a later stage (e.g. Toxvaerd (2006)).

R&D projects are furthermore often embedded in larger contexts which impose deadlines for their completion. Reasons for deadlines can be seen in the (likely) date by which a competitor will have solved a similar problem, or in the limited availability of resources (laboratories, scientific equipment). In our stylized model, we restrict the timescale to two periods and the deadline is at the end of period 2.

In addition, researchers typically exhibit a high degree of specific knowledge, making their actions and results hard to evaluate for any less knowledgeable party. Therefore, a (less informed) principal must to some extent rely on the agent's reports on project status. An asymmetry in observability is plausible: While a researcher will generally be unable to pretend that an unfinished project has been completed, s/he may very well be able to conceal its completion. This limited observability of the completion status gives rise to the principal's problem of discriminating between different types of agents, successful and unsuccessful ones after period 1, and a rational principal would already have to consider the agent's potentially untruthful reporting at the contracting stage. This is a typical problem of hidden information. Unlike the canonical model, however, in our model contracting takes place before the information asymmetry comes about, and the type distribution is *endogenously* determined, viz. by the effort choice for period 1, which in turn depends on the working incentives that are induced by the contract. The problems of hidden actions and hidden information are thus closely linked in our model.

For certain parameter constellations, specifically if early completion of the project is particularly desirable, strategic delay is neither a problem nor an observable phenomenon. The optimal incentive-compatible contract ensures the truthful revelation of the project status as a by-product. However, if obtaining a solution late rather than never is the driving interest, then the truth-telling constraint binds and the first best is not obtainable. The principal constructs the contract so as to deter the agent's strategic delay of a success report. In doing so, the agent's effort in the two periods can no longer be separately incentivized. As a consequence, either the agent's efforts deviate from his or her first-best levels in both periods or the principal strategically delays the start of the project, making

early success impossible. Thus, strategic delay on the part of the agent is a problem, but not a phenomenon on the equilibrium path. In fact, if the principal does not strategically delay the start of the contract, early success will be more likely than in the first-best, due to the distortion of the agent's incentives. If this distortion is too severe, the principal will prefer to strategically delay the start of the project and the optimal contract induces first-best effort in the second period. In our model setting, if a strategic delay occurs, it is due to not allowing work to fill the time that would in principle be available.
The remainder of the paper is structured as follows: Section 2.2 relates our research to the literature. In Section 2.3, we present our model, derive the circumstances under which the first best is (not) attainable and characterize the optimal contract. We also discuss further contractual frictions that may be relevant in the context of our model, in particular limited liability and limited commitment power on the part of the principal. In Section 2.4, we extend the scope of our model. We discuss limits to the principal's commitment power, and possible institutional remedies to the problem of strategic delay, in particular the possibility of contracting with more than one agent. We conclude in Section 2.5. For proofs and mathematical details of our arguments, we refer to the Appendices.

2.2 Related Literature

Our work is related to various strands of the existing literature. Most broadly, our work contributes to the literature on (dynamic) agency, in which a principal wishes to set proper working incentives such that the agent's (intertemporal) performance maximizes the principal's benefit, e.g. Holmström (1979), Lambert (1983), Holmström and Milgrom (1987) and Malcomson and Spinnewyn (1988). More recent works in that strand of literature that - unlike the present paper - use continuous time frameworks comprise, among others, Sannikov (2008), Biais et al. (2007, 2010), Hoffmann and Pfeil (2010) and Williams (2015). None of these papers explicitly address the problem of deadlines; rather, the focus of these papers is on analyzing the agent's incentives to divert cash flows for private benefit and what contractual countermeasures can be taken by the principal.
There is also a rich body of literature that explicitly analyzes R&D settings from an agency perspective (e.g. Manso (2011)). In many of these contributions, the parties involved learn about the project value over time and make the continuation of the project dependent on intermediate project outputs (e.g. Bergemann and Hege (1998, 2005),

Hörner and Samuelson (2013)). Typically, the informational friction between principal and agent combined with the project's unknown returns results in the funding of research projects being stopped too early. Unlike these works, we do not consider the case of an uncertain or steadily updated project value in our own model.

Instead, we focus on the information asymmetry that arises between principal and agent because of the unobservability of the project's completion status. There are a few contributions that have chosen a similar approach. Most notably, Green and Taylor (2016) analyze a two-stage project setting where the project's progress level is visible only to the agent and the agent's self-reported project status is directly contractible. The optimal incentives scheme uses the potential termination of the research project as an incentive device to prevent both shirking and making false statements about the project's actual progress level. In a similar fashion, Lewis and Ottaviani (2008), Lewis (2012) and Ulbricht (2016) analyze models of delegated search where the agent is able to underreport the obtained results. All three contributions substantially differ from ours in numerous ways. In Lewis and Ottaviani (2008), the search revenues are taken from a continuous distribution and are decreasing over time, rendering the optimal speed of search the central question of the analysis. In the paper of Lewis (2012), a search for the best alternative is analyzed. Once more, deadlines appear endogenously from the principal's quest to not fund an agent who has already made a valuable discovery. Ulbricht (2016) focusses on a delegated search where the distribution of search revenues is unknown, but can be disclosed by the agent. In contrast to the works cited above, our paper uses a discrete two-time-period setting, we do not impose wealth constraints on the agent, and, most importantly, the possibilities to choose the contract duration are limited due to the presence of an exogenous deadline in our model. Campbell et al. (2014) present an insightful model in which multiple agents can complete a joint project and might withhold output from each other (but not from a principal, as in our work).

Moreover, our paper is related to further contract-theory contributions that deal with incentives for project completion or potential delays. Mason and Välimäki (2015) study optimal contracts for project completion in a continuous time setting and analyze the resulting optimal contracts with and without the principal's ability to commit to a long-term contract. Toxvaerd (2006) models the execution of a multistage project under agency that requires multiple milestones to be achieved for its completion. Due to the agent's

risk aversion the per-period effort level is generally lower than in a first-best setting, resulting in longer completion times as compared to a non-delegated project. Katolnik and Schöndube (2019) present a model in which the agent derives a private benefit from holding the position of project manager and therefore has an incentive to delay the project's completion. Models that analyze the problem of delayed output as a result of an agent's time-inconsistent preferences are provided, for example, by O'Donoghue and Rabin (1999) and Herweg and Müller (2011).

2.3 The Model

A risk-neutral (female) principal seeks to complete a research project before its deadline, which is two periods ahead. She can delegate the research to a risk-neutral (male) agent, whose research efforts are not observable. The principal's preferred effort choices are our benchmark for the first best. Thus, we abstract from the possibility that the completion of the research project has a social value beyond the principal's benefits. The unobservable effort $e_t \in [0, 1]$ in period $t = 1, 2$ determines the probability $\rho(e_t)$ of successfully completing the project in period t, $\rho(e_t) = e_t$. The project can either be successfully completed early ($t = 1$), late ($t = 2$) or never (ultimate project failure). Research effort comes at a strictly convex cost $C(e_t)$, with $C(0) = C'(0) = 0$.

Project success is observable only to the researcher (i.e. to the agent), and it is verifiable. A failure is not verifiable, but payments can condition on a failure to report a success. The principal's payoff from the project depends on the project status and the timing of its revelation, as depicted in Table 2.1.

The principal obtains a payoff $Y_E > 0$ if an early success (i.e. in period 1) is reported, and $Y_L > 0$ if a success is achieved and reported in period 2. If the project ultimately ends in failure or a success is never reported, the principal obtains $Y_N < Y_E, Y_L$. If the agent withholds an early success and reports a successful completion of the project only in period 2, the principal obtains Y_D. While the principal realizes the payoff consequence of this strategic delay, it is not verifiable whether the success has been achieved early or late.

Outcome in $t = 1$	Success	Failure
Report in $t = 1$	Success	Failure
Payoff in $t = 1$	Y_E	0

Outcome in $t = 2$	-		Success	Ultimate Failure
Report in $t = 2$	-	Success	Success	Ultimate Failure
Payoff in $t = 2$	-	Y_D	Y_L	Y_N

Table 2.1: The principal's payoffs as a function of outcomes and reports in $t = 1, 2$

The principal can specify three payments for the verifiable events, w_E, w_L, w_N, i.e. the agent earns a wage w_E for reporting a success in period 1, w_L for reporting a success in period 2, and w_N in case he fails to report a success in either period.
We assume that the principal and agent are rational, and that they maximize their expected discounted payoff (net of the cost of effort). We assume a common discount factor δ and that our parameter constellation satisfies:

A1 $Y_E - \delta \cdot Y_N < C'(1), Y_L - Y_N < C'(1)$.

A2 $Y_E > \delta \cdot Y_D$.

We impose these assumptions in order to avoid case distinctions in our statements. Assumption A1 rules out that maximal effort provision (and hence certain completion of the project) is optimal in any period. We choose a parsimonious model representation based on this parameter constellation, remaining agnostic about its underlying reasons, as there may be many factors that impact on it. For instance, a successful project completion may induce a constant stream of payoffs starting at the moment of completion, in which case $Y_E > Y_L$, or it may give rise to a payoff (e.g. attention as the main source of the principal's benefit) only in the period of its publication, in which case Y_L could be greater than Y_E. Ultimate project failure may be more severe than a (preliminary) failure to achieve a success in the first period ($Y_N < 0$), or it may be equally bad. We assume that a strategically delayed success report harms the principal in order to rule out a (trivial) preference-based explanation for strategic delay (Assumption A2). If instead $\delta \cdot Y_D \geq Y_E$, the principal actually prefers the agent to strategically delay the report of an early success (maybe because she herself only receives research funds for as long as the project remains incomplete). In this case, strategic delay would be a phenomenon that occurs due to the principal's preferences, but it would not be a problem. Several reasons come to mind

why a strategic delay can be harmful. The research result may leak, the exploitability of the discovery may be limited, or the principal may have to bear costs for continuing the research in the second period. Moreover, the priority principle in science contributes to the desire to avoid a delay in publishing a research result.

2.3.1 First Best

For now, we will neglect the agency problems and consider the principal's optimization problem as though she were able to control the effort levels directly (e.g., performing the research herself). We refer to the solution that maximizes the principal's payoff as first best. The principal's ex ante objective function can then be written as:

$$V_1 = e_1 \cdot Y_E - C(e_1) + (1 - e_1) \cdot \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - C(e_2)) . \tag{2.1}$$

If the project was completed in the first period, no further research is carried out. If the project was not yet completed in the first period, the optimal effort in the second period maximizes:

$$V_2 = Y_N + e_2 \cdot (Y_L - Y_N) - C(e_2), \tag{2.2}$$

where $C'(0) = 0$ makes sure that the project is worth continuing in $t = 2$ and rules out the corner solution $e_2 = 0$. The first order condition reads:

$$Y_L - Y_N = C'(e_2). \tag{2.3}$$

As the marginal benefit is constant and the marginal cost increases in e_2, problem (2.2) has a unique solution. The corner solution $e_2 = 1$ is ruled out by Assumption A1. Consequently, (2.3) defines the optimal level e_2^*.[9] Denote with V_2^* the (optimized) expected value of the second period research, given a failure in the first period (evaluated in the second period):

$$V_2^* = Y_N + e_2^* \cdot (Y_L - Y_N) - C(e_2^*), \tag{2.4}$$

where e_2^* solves (2.3). Note that e_2^* is independent of the effort level chosen in the first period. The principal chooses e_1 so as to maximize:

$$V_1 = e_1 \cdot (Y_E - \delta \cdot V_2^*) - C(e_1), \tag{2.5}$$

[9] The second order conditions are presented in the proof of Proposition 2.1 in Appendix A.

where V_2^* is defined in (2.4) and e_2^* satisfies (2.3). The condition for an interior optimum reads:

$$Y_E - \delta \cdot V_2^* = C'(e_1). \tag{2.6}$$

If equation (2.6) has a solution, it is unique, as the marginal cost of effort is strictly increasing and the marginal benefit is constant. The corner solution $e_1^* = 1$ is ruled out by Assumption A1.
We find that:

Proposition 2.1. *The first-best combination of efforts e_1^*, e_2^* is unique. It is optimal to report an early success in the first period.*
(i) If $Y_E > \delta \cdot V_2^*$, where V_2^* is defined in (2.4), $e_1^*, e_2^* \in (0,1)$, e_1^* is defined by (2.6) and e_2^* is defined by (2.3).
(ii) If $\delta \cdot V_2^* > Y_E, e_1^* = 0, e_2^* \in (0,1)$, and e_2^* is defined by (2.3).

Proof: See Appendix B.1.

The optimal reporting behavior follows from Assumption A2. If the LHS of (2.6) is negative (case (ii) in the above Proposition), the principal prefers not to provide any effort at all in the first period at all, i.e. $e_1^* = 0$. In the following, we assume that the parameter constellation is such that the LHS of (2.6) is strictly positive, such that it is optimal to induce some effort in period 1. Otherwise, if no effort is optimally provided in period 1, we obtain another preference-based explanation for a delay of the project, and the strategic delay of a success report is not an issue.
The benefit of a marginal unit of effort is $Y_E - \delta \cdot V_2^*$ in the first period, and $Y_L - Y_N$ in the second period. The component $\delta \cdot V_2^*$ represents the expected payoff from continuing the research project in the second period, which is forgone if the project is completed early. When $\delta \cdot V_2^*$ exceeds $Y_E - (Y_L - Y_N)$, the benefit from providing a marginal unit of effort early rather than late, it follows from the first order conditions that $e_2^* > e_1^*$. In such parameter constellations, it is ex ante more likely that the project will be completed close to the deadline than ahead of the deadline due to the optimal choice of efforts. A failure in $t = 1$ tends to be less harmful than a failure in $t = 2$, since a successful completion of the research project is still possible after the former, but not after the latter event.

2.3.2 Delegated Research

We now turn to the analysis of our agency problems. We intend to identify the circumstances under which the incentives for efficient effort provision are (in)compatible with the incentives for truthfully revealing an early success. Remember that a success can only be reported if the project has indeed been successfully completed. If the agent reports an early success, he is rewarded with w_E. If the agent delays the report of a success until the second period, his reward is w_L in the second period, the same as if he reports a success achieved in period 2. If the agent fails to report a success in both periods, he obtains w_N in the second period. The agent's ex ante participation constraint for $t = 1$ is therefore:

$$R_1 = e_1 \cdot (w_E - \delta \cdot w_N) - C(e_1) + (1 - e_1) \cdot \delta \cdot (e_2 \cdot (w_L - w_N) - C(e_2)) + \delta \cdot w_N \geq 0. \quad (2.7)$$

The agent acts rationally and anticipates his choices in the second period. If the project has not been completed in the first period, he chooses the effort in the second period so as to maximize his expected payoff in that period. In order to induce the efficient choice of effort in the second period, the principal needs to align the agent's interests with her own. Consequently, by using the first order approach and making use of (2.3), the incentive constraint for efficient effort provision in the second period reads:

$$w_L - w_N = Y_L - Y_N. \quad (2.8)$$

Denote with R_2 the expected rent (possibly) granted to the agent in period 2:

$$R_2 = e_2 \cdot (w_L - w_N) - C(e_2) + w_N, \quad (2.9)$$

with $e_2 = e_2^*$ if (2.8) holds.
In order to align interests in the first period, making use of (2.6), (2.8) and (2.9), the principal has to set the reward for a success reported in the first period as follows:

$$w_E = Y_E - \delta \cdot (V_2^* - R_2), \quad (2.10)$$

where $V_2^* - R_2 = Y_N - w_N$ if (2.8) holds, such that:

$$w_E = Y_E - \delta \cdot (Y_N - w_N). \quad (2.11)$$

In the absence of the strategic delay problem, the agent's choice of efforts coincides with the first-best solution if and only if (2.8) and (2.11) both hold. The principal can then satisfy the agent's ex ante participation constraint by choosing w_N to satisfy:

$$e_1^* \cdot w_E - C(e_1^*) + (1 - e_1^*) \cdot \delta \cdot (e_2^* \cdot w_L + (1 - e_2^*) \cdot w_N - C(e_2^*)) \geq 0. \tag{2.12}$$

No loss of efficiency is triggered by the delegation of the choice of efforts alone, as (2.8), (2.11) and (2.12) can be satisfied simultaneously.

However, the agent's discretion regarding the choice of whether and when to report a success imposes additional constraints. In the second period, truthful reporting is a by-product of incentive provision, as $w_L > w_N$. A problem may arise in the first period, as the agent may profit from a strategic delay of his report of a success. Truth-telling requires:

$$w_E \geq \delta \cdot w_L. \tag{2.13}$$

The conditions for efficient effort provision are compatible with truth-telling if and only if:

$$Y_E - \delta \cdot (Y_N - w_N) \geq \delta \cdot (Y_L - Y_N + w_N), \tag{2.14}$$

that is, if and only if:

$$Y_E \geq \delta \cdot Y_L. \tag{2.15}$$

If (2.15) is violated, the agent strategically withholds the report of a success in the first period. Anticipating the strategic delay, it is no longer (2.11) which guides the agent's effort choice in the first period, because he does not plan to cash in w_E, but $\delta \cdot w_L$ for an early success. We conclude:

Proposition 2.2. *If the principal delegates research to the agent, the effects of an unobservable project completion status are as follows:*

(i) If $Y_E \geq \delta \cdot Y_L$, the first best is implementable.

(ii) If $Y_E < \delta \cdot Y_L$, the first best is not implementable. If the principal naïvely offers the contract that is optimal in the absence of the strategic delay problem, the agent's effort choice is inefficiently high in the first period and he strategically delays the report of an early success.

If $Y_E \geq \delta \cdot Y_L$, the incentive to report truthfully is implied by the incentives that induce efficient effort provision. Strategic delay is neither a phenomenon nor a problem (in our otherwise agency-friendly model). The agent's working incentives can be set such that the agent effectively internalizes the principal's payoff consequences when choosing his effort levels. The principal can choose the agent's payoffs so as to maximize the surplus by satisfying (2.8) and (2.11), and she can extract the entire surplus by satisfying (2.12) with equality:

$$w_N = -\left(\frac{1}{\delta} \cdot (e_1^* \cdot Y_E - C(e_1^*)) + (1 - e_1^*) \cdot (e_2^* \cdot (Y_L - Y_N) - C(e_2^*))\right). \tag{2.16}$$

Strategic delay (possibly) occurs only if $\delta \cdot Y_L > Y_E$. In this case, the incentives for efficient effort provision are not compatible with the incentives to immediately report an early success. Then, strategic delay is a problem (because we assumed that $\delta \cdot Y_D < Y_E$), and, if unaccounted for, it is certainly also a phenomenon. If strategic delay on the part of the agent is deterred in the optimal contract (and is hence not a relevant phenomenon on the equilibrium path), this deterrence comes at the cost of distorting the incentives for effort provision.

The good news from our analysis so far is that for a range of plausible parameter constellations, strategic delay does not appear to be a problem. However, this finding should be interpreted cautiously given our choice of a particularly agency-friendly model framework (no risk aversion, no limits to liability). In our model, the principal can extract the entire surplus while aligning the agent's payoff consequences of his actions with her own, except for those of a delayed report. In a less agency-friendly setting, the problem of strategic delay will certainly interact with other sources of distortion.

In the next section, we show how the principal addresses the problem of strategic delay in her optimal contract with the agent.

2.3.3 Optimal Contracting Under Strategic Delay

In this section, we focus on the parameter constellation $Y_E < \delta \cdot Y_L$ in order to study optimal contracting when strategic delay is a problem. For the complementary case, the optimal contract is defined by the incentive-compatibility constraints (2.8) and (2.11) and the participation constraint (2.12).

The principal has four options: She could allow the agent to delay the report of an early success, deliberately violating the truth-telling constraint (2.13) by offering a *pooling contract*. With such a contract, the project's completion, if achieved, would be revealed in $t = 2$ in any case. She could ensure the agent truthfully reports a success in the first period by designing the contract as a *separating contract*. She could strategically *delay the start of the project* by offering a one-shot contract for $t = 2$. Lastly, she could *terminate* the contractual relationship *early* at the end of the first period. We will analyze these four options in turn.
We start by showing that a pooling contract can never be optimal.

Proposition 2.3. *A pooling contract is never optimal.*

Proof: Suppose the principal offers some pooling contract w_E, w_L, w_N, where (2.13) is violated such that w_E is too low to be relevant on the equilibrium path. The contract induces certain effort levels on the part of the agent, and the late reporting of an early success, in which case the principal would earn $\delta \cdot Y_D$. If the principal offered a wage $w_E = \delta \cdot w_L$ instead, the agent would be willing to report an early success in the first period, in which case the principal earned $Y_E > \delta \cdot Y_D$. As $\delta \cdot w_L$ is in any case the relevant incentive to provide effort in the first period, the agent's effort choices would not be affected by such a modification of the contract, nor would his incentive to participate in the contract. Hence, the proposed modification of the contract would leave the principal better off and the agent as well off as under the original (pooling) contract. It follows that the pooling contract is not optimal. □

Next, we turn to the optimal separating contract. We use the superscript S for the endogenous variables and characterize the optimal separating contract as follows:

Proposition 2.4. *Suppose $Y_E < \delta \cdot Y_L$. Compared to the first-best, the optimal separating contract induces effort levels such that $e_1^S > e_1^*$ and $e_2^S < e_2^*$.*

In Appendix B.1, we provide a complete derivation of the optimal separating contract. We demonstrate that it is not optimal for the principal to leave either the participation constraint or the truth-telling constraint slack. The binding truth-telling constraint $w_E^S = \delta \cdot w_L^S$ effectively leaves two choice variables for the principal, a payment in the case of

a successful completion of the research project (appropriately discounted if it is reported early), and a (negative) payment in the case of an ultimate failure. The principal chooses the spread to guide the agent's incentives, and she uses the failure payment in order to satisfy the participation constraint. She therefore cannot target the agent's effort choices in both periods separately, which leads to the compromise in between the optimal levels, as stated in the proposition. It is interesting to note that our model predicts a *higher* probability of observing an early success in delegated research over the two periods than if the principal carries out the research on her own.

Next, we address the principal's third option, i.e. strategically delaying the start of the project. Not being exposed to the possibility of doing research in the first period, the agent cannot hide an early success, making the truth-telling constraint irrelevant. Due to the convex cost of effort, the agent has an interest in smoothing his effort. Thus, if the agent anticipates that a contract will be offered to him in the second period, effort-smoothing can only be prevented if the principal can effectively exclude the agent from the research technology. She has an interest to do so if Y_E is sufficiently small:

Proposition 2.5. *Consider $Y_E < \delta \cdot Y_L$. There is a threshold $\overline{Y}_E > \delta \cdot Y_L$ such that for $Y_E < \overline{Y}_E$, the principal prefers to strategically delay the start of the project until period 2. For $Y_E > \overline{Y}_E$, she prefers to offer the optimal separating contract.*

Proof: Remember Proposition 2.1(ii): For $Y_E = \delta \cdot V_2^*$ we have $e_1^* = 0$, such that for this parameter range, the proposition is in fact a straightforward corollary to Proposition 2.1(ii). Consider the case $Y_E = \delta \cdot V_2^* + \epsilon$, with $\epsilon > 0$. For ϵ close to zero, e_1^* is close to zero and so is the first period's contribution to the principal's ex ante expected payoff. The spread in the optimal separating contract $w_L^S - w_N^S = \Delta$ induces an effort level in the first period that is discretely higher than e_1^* and induces an effort level in the second period that is strictly below e_2^*, such that the principal's expected payoffs are lower than first best in both periods. In fact, the deviation of Δ from $Y_L - Y_N$, (the level that induces $e_2 = e_2^*$), is mainly due to preventing an excessively high effort level in period 1. Thus, for ϵ close to zero, the principal prefers to enforce $e_1 = 0$ without distorting the second period incentives. This proves the first part of the proposition.

At the other end of the relevant parameter range, i.e. for $Y_E = \delta \cdot Y_L - \epsilon$, ϵ close to zero, the spreads that induce the first-best levels of effort in period 1 and 2, respectively, differ

only marginally. Respecting the truth-telling constraint induces only small distortions of the agent's incentives from their first-best levels in the optimal separating contract. Moreover, the principal's expected payoffs in both periods deviate only marginally from their first-best levels, and they contribute almost equally to her overall expected payoff. Hence, the principal strictly prefers the separating contract over a delay of a project start. The existence of a (unique) threshold follows from the fact that the principal's payoff is strictly increasing in Y_E in the separating contract, and is not affected by it when delaying the start of the project. □

Finally, we consider the principal's last option, i.e. concluding a contract which incentivizes effort only in period 1, which would render the strategic delay problem obsolete.

Proposition 2.6. *Consider $Y_E < \delta \cdot Y_L$. A contract that deters research in period 2 is dominated by a separating contract.*

Proof: If the principal does not incentivize effort in the second period, her payoff in the second period (conditional on a failure in the first period) is Y_N, the lowest possible value for V_2. Denote the payments to the agent in the first period conditional on a reported success and failure, respectively, with $\tilde{w}_E$ and $\tilde{w}_N$. The optimal contract in this class of contracts satisfies the agent's participation constraint with equality, and the optimal spread $\tilde{w}_E - \tilde{w}_N$ equals $Y_E - \delta \cdot Y_N$.
Suppose the principal expands this contract to include payments to the agent $\tilde{w}_L, \tilde{w}'_N$ for a success and a failure in the second period, respectively, as follows: $\tilde{w}_L - \tilde{w}'_N = \epsilon > 0$, $\epsilon < \min\{Y_E, \frac{\tilde{w}_E}{\delta}\}$, $\tilde{w}_L \leq \frac{\tilde{w}_E}{\delta}$ (such that the truth-telling constraint is satisfied), and $C'(\epsilon) \cdot \epsilon + \tilde{w}'_N = C(\epsilon)$ (such that the agent's expected rent in the second period is zero). It is easy to verify that such payments exist: Choose any $\epsilon < \min\{Y_E, \frac{\tilde{w}_E}{\delta}\}$ and set $\tilde{w}'_N = C(\epsilon) - C'(\epsilon) \cdot \epsilon$, and $\tilde{w}_L = \epsilon + \tilde{w}'_N$. As $\tilde{w}'_N < 0$ (due to the convexity of $C(\cdot)$), $\tilde{w}_L < \frac{\tilde{w}_E}{\delta}$.
These modifications to the original contract mean that neither the agent's participation constraint nor his working incentives in period 1 are affected. The agent provides a positive level of effort in period 2, $\tilde{e}_2 = \epsilon < Y_E < \delta \cdot (Y_L - Y_N) < Y_L - Y_N$, which gives rise to an expected payoff for the principal strictly between Y_N and V_2^* in period 2 (conditional on a failure in period 1).

The principal's expected payoff in the first period is the same as in the original contract, whereas the expected payoff in the second period is strictly higher. Thus, she is better off under the modified contract than under the original contract. □

To summarize the results of our analysis: The principal deals with the strategic delay problem either by offering a separating contract, or by strategically delaying the start of the project. The separating contract distorts the agent's efforts in both periods, and leads to a strictly higher probability of solving the research problem in the first period. A delay of the start of the project allows the first-best effort to be implemented in the second period, but the opportunity to find a solution in the first period is completely forgone.
The latter case is somewhat reminiscent of the work of Bhaskar (2014), who presents a long-term agency model in which the principal learns about the agent's ability over time and an inefficient zero-effort in early periods can be optimal.

2.4 Model extensions

In this section, we intend to explore the strategic delay problem in a richer institutional setting. We comment on the limits to our model, and we augment our model in several directions.

2.4.1 Limits to Enforceability

So far, we have assumed that the principal is able to enforce the ex ante optimal contract. This assumption is crucial for the viability of both solutions to the strategic delay problem, the separating contract and the strategic delay of the project start. The optimal separating contract induces an inefficiently low effort level in period 2, which gives rise to scope for renegotiation and challenges the principal's commitment power.
Likewise, whether the principal is able to delay the project start and effectively prevent early research depends on whether the agent can anticipate the contract to be concluded in the second period. If he can do that, he is better off smoothing the cost of effort, and he would already start the research (provided he has access to the appropriate research technology) in period 1. In such a case, a strategic delay of the project start is de facto impossible and the principal would be better off offering the optimal separating contract.
If the principal cannot commit not to renegotiate a separating contract upon a failure

in period 1, she will optimally propose a renegotiation-proof contract in the first place, which implements first-best incentives for the second period.[10] This fact, the binding truth-telling constraint and the binding participation constraint jointly pin down the optimal separating contract, which induces an even higher effort level in period 1 than the optimal separating contract in the case with commitment. This distortion further decreases the principal's ex ante expected payoff and makes the separating contract less attractive. As a consequence, the parameter range for which a strategic delay of the project start is optimal increases. An analogous reasoning to the proof of Proposition 2.5 implies that there remains a parameter range for which the optimal solution to the strategic delay problem is a separating contract.

2.4.2 Multiple Agents

In many real world settings it is plausible to assume that there will be more than one agent working to achieve a particular research goal. In this section, we analyze whether the principal can profit from inducing competition among multiple agents. Making any agent's payment contingent on that agent's individual report *and* also on the report of other agents seems like a promising way to mitigate, or sidestep altogether, the problem of strategic delay. In this section, we outline the consequences of introducing a second agent, such that in total there are two agents, named A and B. We refer to Appendix B.2 for a detailed derivation of the results presented here.

In what follows, we stipulate that the parameter constellation is such that for both agents the respective effort levels yield interior solutions which are period-wise symmetric, i.e. $e_t^A = e_t^B$ for $t = 1, 2$. Let $P(Y_E) = (1 - (1 - e_1^A) \cdot (1 - e_1^B))$ denote the probability of an early success. Likewise, we define $P(Y_L) = (1 - (1 - e_2^A) \cdot (1 - e_2^B))$. The principal's maximization problem now reads as:

$$P(Y_E) \cdot Y_E - 2 \cdot C(e_1^A) + (1 - P(Y_E)) \cdot \delta \cdot (Y_N + P(Y_L) \cdot (Y_L - Y_N)) - 2 \cdot C(e_2^A)). \quad (2.17)$$

Due to the symmetry, it suffices to express the solution to the principal's problem in terms of the effort levels demanded from agent A (analogous conditions with the superscripts reversed apply to agent B's effort levels).

[10] For an in-depth analysis of the principal's ability to commit to long-term contracts and its role in project completion, see Toxvaerd (2006).

Optimal effort levels are implicitly defined by:

$$C'(e_1^A) = (1 - e_1^B) \cdot (Y_E - \delta \cdot V_2^{**}), \tag{2.18}$$

$$C'(e_2^A) = (1 - e_2^B) \cdot (Y_L - Y_N), \tag{2.19}$$

where $V_2^{**} \geq V_2^*$, because the expected profit attainable with two agents is generally higher than with one agent. Due to the introduction of a second agent, the contracting space has become larger. In particular, we can condition any agent's compensation level on the reports obtained from both agents. Therefore, we analyze a scenario in which the agent that is the only agent to report an error after period 1, has to pay a penalty w_F. Apart from that, all other payments remain defined as in the single-agent case.
Then, agent A chooses e_1^A so as to maximize:

$$e_1^A \cdot w_E + (1 - e_1^A) \cdot (e_1^B \cdot w_F + (1 - e_1^B) \cdot \delta \cdot R_2) - C(e_1^A). \tag{2.20}$$

The adjusted truth-telling constraint reads as:

$$w_E \geq e_1^B \cdot w_F + (1 - e_1^B) \cdot \delta \cdot w_L. \tag{2.21}$$

In Appendix B.2 we show that condition (2.21) can be rearranged to give:

$$Y_E \geq \delta \cdot (Y_L - C(e_2^A)). \tag{2.22}$$

Comparing (2.22) with (2.15), we observe that the parameter range for which truth-telling is incentive-compatible with first-best incentives has increased. Interestingly, it is not the punishment but the presence of the second agent alone that causes this parameter shift.[11] In fact, w_F can be chosen arbitrarily. In order to guide first period incentives for effort provision effectively, w_E has to be correspondingly increased if we impose a punishment. As a consequence, the punishment is canceled out when it comes to truth-telling.
The fact that strategic delay does indeed help reduce the parameter range (as compared to the single-agent case) for which strategic delay is relevant follows from the interplay of both agents' incentive compatibility constraints that make the agent on the margin the

[11] It is beyond the scope of this work to fully analyze the complete set of possible payments as a function of the agents' reports. Still, our result is indicative that punishments are not conducive to mitigating the problem of strategic delay in settings with multiple agents.

residual claimant to the social consequences of his actions. If an agent reports truthfully in the first period, the principal gets Y_E instead of entering the next period. If instead she enters the next period, the other agent will be present as well. Taking that as a given, the prospective earning is only $Y_L - C(e_2^A)$ (instead of Y_L in the single-agent case), which causes the truth-telling condition to be adjusted.

2.4.3 Replacement of Agents

If strategic delay poses a problem, i.e. it reduces the principal's expected gain compared to the first-best benchmark, a further obvious solution to the problem would be to conclude two separate short-term contracts with two distinct agents. The first-best effort levels would be implementable for the principal in two separate one-shot contracts, one contract for agent 1 in $t = 1$ and a separate one for agent 2 in $t = 2$, which is offered to a second agent only in case of a failure in the first period.[12] Truthful reporting of a project success would then not pose a problem in either period, since in a one-shot contract, truth-telling is a by-product of efficient effort provision. However, one obvious objection that could be raised here is that a series of short-term contracts might cause extra costs compared to one single long-term contract, e.g. due to substantial transaction costs associated with searching for and hiring agents.[13] Secondly, it can be assumed that non-negligible costs are also associated with the (initial) training of the new agent. Formally, this means that $C(0) > 0$ in $t = 1$ and also in $t = 2$, provided that the agent is replaced after the first period. These additional costs might outweigh the loss caused by the information asymmetry which only accrues in the single-agent case.

2.4.4 Monitoring

A further coping strategy from the principal's perspective is to actively reduce the information asymmetry between herself and the agent. Assume that the principal can invest some amount k in costly monitoring, such that the actual project completion status after period 1 becomes observable to her with certainty (and verifiable to any third party).[14] The contract could then include a punishment, i.e. a payment $P < 0$, if the agent's

[12] We need to exclude the possibility of side-contracting between agent 1 and 2.

[13] This argument is related to Coase's rationale for the superiority of hierarchies over markets (Coase (1937)).

[14] Related ideas can be found in the literature on costly state verification, e.g. Townsend (1979). There is also a section on monitored search in Lewis and Ottaviani (2008).

strategic delay is discovered (a less formal means of punishment would be to expose his misconduct to the scientific community). If the principal can commit to engage in costly verification of a reported failure in period 1 with some probability q, the first-best solution would (almost) be achievable for the principal, as long as there is no upper limit to the amount of the punishment. This is the case because the principal will choose q at just high enough a level to satisfy the (modified) truth-telling constraint $w_E \geq \delta \cdot w_L + q \cdot P$. The harsher the punishment becomes, the closer to zero q can be set. Hence, the principal has (virtually) no monitoring costs, while the agent is still willing to accept the contract. On the equilibrium path, he will never choose not to report an early success and is therefore never punished. Potential upper limits to the amount of the punishment, the principal's inability to commit and the difficulty to prove an unreported success in court constrain this idea's viability in practice. If the principal cannot commit to an ex ante probability of monitoring, the possibility of monitoring and punishing gives rise to an inspection game with an equilibrium in mixed strategies (see Chapter 3).

2.4.5 Plurality of Research Methods

In many research settings a specific research goal (the development of a new vaccine, say) can be achieved using one of a number of methods or technologies. The choice of technology, then, could also be harnessed to prevent strategic delay if we assume that the technology used to find a solution to the research problem can be inferred from the solution, and that this information is contractible and observable to parties other than the agent. Then, the choice of research technology could be made an explicit part of the contract. Suppose, for example, that in $t = 1$ research is to be carried out with technology a and, upon failure, with research technology b in $t = 2$. Hence, early completion with technology a cannot be incorrectly presented as late completion if the contract specifies that technology b should have been used in the second period and that the incorrect choice of technology can be deterred by inflicting sufficiently severe punishments on the agents. If technology b is potentially less promising or powerful than technology a, the principal faces a trade-off between the advantage of not being confronted with the problem of strategic delay and the disadvantage of having the research performed with an inferior technology (see also Chapter 1 where we present an in-depth study of the problem of technology choice).

2.5 Discussion and Conclusion

We have provided an (admittedly stylized and parsimonious) agency model that analyzes the tension between providing working incentives on the one hand and incentives to truthfully reveal information on the other in the context of R&D projects. We have identified conditions where this tension actually exists. If the contract offered does not take into account the possibility that the agent might withhold information on an early success despite being incentivized to do so are induced by a naïvely concluded contract, the agent will "overinvest but underreport" in $t = 1$, i.e. he will put in more effort than the principal actually desires but delay the disclosure of an early success. The principal's rational adjustment of the contract hinges on her possibility to effectively prevent the agent from conducting research in the first period. If this is not feasible, a separating contract that disincentivizes strategic delay but comes at the cost of distorted effort levels is the principal's best option. If a strategic delay of the project start is feasible, the principal will prefer this option if her payoff from an early success is sufficiently low.

The good news from our analysis is that parameter regions exist where truthful reporting is a by-product of incentive provision, in which case strategic delay is neither a problem nor a phenomenon. If the principal does not suffer from strategic delay (i.e. if our Assumption A2 does not apply), strategic delay will be a phenomenon if and only if it is not a problem. These results should however be interpreted with caution, as we study a setting with otherwise ideal contracting conditions (e.g. no risk aversion, no constraints on payments). If the principal cannot fully extract the surplus due to other contractual frictions, the problem of strategic delay is likely to interact with the distortions from such frictions.

Possible extensions and avenues for further research are an analysis of the agent's risk preferences or consideration of more than only two periods. Furthermore, specifications of the researcher's yield and cost functions could be analyzed. In our current setting, we do not consider learning effects, such that the effort in $t = 1$ has no effect at all on the probability of generating a success in $t = 2$ (or on the cost of conducting research in that period). While research efforts that lead to an initial failure can often be considered sunk costs, a more flexible representation has the potential to enrich our analysis.

B Appendices

B.1 Proofs

Proof of Proposition 2.1

The candidates for the optimum have been derived in the text. It remains to be shown that the second order conditions are satisfied. We have:

$$\begin{aligned}\frac{\partial^2 V_1}{\partial e_1^2} &= -C''(e_1)\\ \frac{\partial^2 V_1}{\partial e_2^2} &= -(1-e_1)\cdot\delta\cdot C''(e_2)\\ \frac{\partial^2 V_1}{\partial e_1 \partial e_2} &= -\delta\cdot(Y_L - C'(e_2))\end{aligned}$$

$\frac{\partial^2 V_1}{\partial e_1^2}$ and $\frac{\partial^2 V_1}{\partial e_2^2}$ are strictly negative, as $C(\cdot)$ is convex. $\frac{\partial^2 V_1}{\partial e_1 \partial e_2}$ is zero when e_2 satisfies the first order condition, such that $Y_L = C'(e_2)$. Thus, $\frac{\partial^2 V_1}{\partial e_1^2}\cdot\frac{\partial^2 V_1}{\partial e_2^2} - \left(\frac{\partial^2 V_1}{\partial e_1 \partial e_2}\right)^2 > 0$, and the conditions referred to in Proposition 2.1 do indeed characterize the maximum of the principal's objective function. □

Proof of Proposition 2.4

We assume that $w_L^S > w_N^S$ such that $e_2^S > 0$ in any separating contract. The complementary case is discussed (and ruled out) in Proposition 2.6. For the sake of readability, we omit the superscript "S" in the following, as there is no scope for ambiguity.
If the agent accepts a contract (w_E, w_L, w_N) that satisfies the truth-telling constraint, he chooses his effort levels e_1, e_2 to maximize:

$$e_1\cdot w_E - C(e_1) + (1-e_1)\cdot\delta\cdot(e_2\cdot(w_L - w_N) - C(e_2) + w_N). \tag{B.1}$$

The agent's effort in the second period is defined by:

$$e_2 = C'^{-1}(w_L - w_N), \tag{B.2}$$

with $\frac{\partial e_2}{\partial w_L} = -\frac{\partial e_2}{\partial w_N} = \frac{1}{C''(w_L - w_N)} > 0$.

Anticipating a truthful report of an early success, effort in the first period satisfies:

$$e_1 = C'^{-1}\left(w_E - \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))\right), \tag{B.3}$$

if the argument of $C'^{-1}(\cdot)$ in the above equation is positive. As $w_E \geq \delta \cdot w_L$, this is the case for all $w_L \geq w_N$. Using (B.2):

$$e_1 = C'^{-1}\left(w_E - \delta \cdot (w_N + (w_L - w_N) \cdot C'^{-1}(w_L - w_N) - C(C'^{-1}(w_L - w_N)))\right), \tag{B.4}$$

with $\frac{\partial e_1}{\partial w_E} = \frac{1}{C''(w_E - \delta \cdot (w_N + (w_L - w_N) \cdot C'^{-1}(w_L - w_N) - C(C'^{-1}(w_L - w_N))))} > 0$, $\frac{\partial e_1}{\partial w_L} = -\delta \cdot C'^{-1}(w_L - w_N) \cdot \frac{\partial e_1}{\partial w_E} = -\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} < 0$ and $\frac{\partial e_1}{\partial w_N} = -\delta \cdot (1 - C'^{-1}(w_L - w_N)) \cdot \frac{\partial e_1}{\partial w_E} = -\delta \cdot (1 - e_2) \cdot \frac{\partial e_1}{\partial w_E} < 0$.

The principal chooses (w_E, w_L, w_N) so as to maximize her expected payoff:

$$e_1 \cdot (Y_E - w_E) + (1 - e_1) \cdot \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))) \tag{B.5}$$

subject to the incentive-compatibility constraints (B.2), (B.4), the agent's participation constraint (B.6) and the truth-telling constraint (B.7):

$$e_1 \cdot w_E - C(e_1) + \delta \cdot (1 - e_1) \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2)) \geq 0, \tag{B.6}$$

$$w_E \geq \delta \cdot w_L. \tag{B.7}$$

Thus, we have two binding constraints and two weak inequalities to satisfy. We use the Karush-Kuhn-Tucker conditions in order to characterize the principal's optimal choice. We seek to minimize:

$$-e_1 \cdot (Y_E - w_E) - (1 - e_1) \cdot \delta \cdot (Y_N + e_2 \cdot Y_L - (w_N + e_2 \cdot (w_L - w_N))) \tag{B.8}$$

satisfying:

$$
\begin{aligned}
e_2 - C'^{-1}(w_L - w_N) &= 0 \quad \text{(B.9)} \\
e_1 - C'^{-1}\left(w_E - \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))\right) &= 0 \quad \text{(B.10)} \\
-e_1 \cdot w_E + C(e_1) - (1 - e_1) \cdot \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2)) &\leq 0 \quad \text{(B.11)} \\
-w_E + \delta \cdot w_L &\leq 0 \quad \text{(B.12)} \\
\mu_1, \mu_2 &\geq 0 \quad \text{(B.13)} \\
\mu_1 \cdot (e_1 \cdot w_E + C(e_1) - (1 - e_1) \cdot \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))) &= 0 \quad \text{(B.14)} \\
\mu_2 \cdot (-w_E + \delta \cdot w_L) &= 0 \quad \text{(B.15)}
\end{aligned}
$$

Conditions (B.9)-(B.12) are the primary feasibility constraints, (B.13) is needed for dual feasibility and (B.14), (B.15) are the conditions for complementary slackness. We use the multipliers λ_1, λ_2 for constraints (B.9), (B.10) and μ_1, μ_2 for (B.11), (B.12), respectively.

(B.14) and (B.15) allow for four cases:

1. $\mu_1 = \mu_2 = 0$,
2. $\mu_1 = 0, \mu_2 \neq 0$, (B.12) is binding,
3. $\mu_1 \neq 0, \mu_2 = 0$, (B.11) is binding,
4. (B.11), (B.12) are both binding.

Case 1 Suppose the solution satisfies the conditions for Case 1 (both inequalities are slack). Then, it needs to satisfy the following Karush-Kuhn-Tucker optimality conditions:

$$
\begin{aligned}
e_1 + \frac{\partial e_1}{\partial w_E} \cdot (-Y_E + w_E + \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) & \\
+\lambda_2 \cdot \underbrace{\left(\frac{\partial e_1}{\partial w_E} - \frac{1}{C''\left(w_E - \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))\right)} \right)}_{=0} &= 0, \quad \text{(B.16)}
\end{aligned}
$$

$$\underbrace{\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E}}_{=-\frac{\partial e_1}{\partial w_L}} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))))$$
$$-(1-e_1) \cdot \delta \left(-e_2 + \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) \right)$$
$$+\lambda_1 \cdot \underbrace{\left(\frac{\partial e_2}{\partial w_L} - \frac{1}{C''(w_L - w_N)} \right)}_{=0}$$
$$+ \lambda_2 \cdot \underbrace{\left(\begin{array}{c} -\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} + \frac{\delta \cdot \left(e_2 + \frac{\partial e_2}{\partial w_L} \cdot \underbrace{((w_L - w_N) - C'(e_2))}_{=0} \right)}{C''\left(w_E - \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))\right)} \\ \\ \end{array} \right)}_{=0} = 0, \quad \text{(B.17)}$$

$$\underbrace{\delta \cdot (1-e_2) \cdot \frac{\partial e_1}{\partial w_E}}_{=-\frac{\partial e_1}{\partial w_N}} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))))$$
$$-(1-e_1) \cdot \delta \cdot \left(\begin{array}{c} -(1-e_2) - \underbrace{\frac{\partial e_2}{\partial w_L}}_{=-\frac{\partial e_2}{\partial w_N}} \cdot (Y_L - Y_N - (w_L - w_N)) \\ \\ \end{array} \right)$$
$$+\lambda_1 \cdot \underbrace{\left(-\frac{\partial e_2}{\partial w_L} + \frac{1}{C''(w_L - w_N)} \right)}_{=0}$$
$$+\lambda_2 \cdot \left(\begin{array}{c} -\delta \cdot (1-e_2) \cdot \frac{\partial e_1}{\partial w_E} + \frac{\delta \cdot \left(1 - e_2 + \frac{\partial e_2}{\partial w_N} \cdot \underbrace{((w_L - w_N) - C'(e_2))}_{=0} \right)}{C''\left(w_E - \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))\right)} \\ \\ \end{array} \right) = 0. \quad \text{(B.18)}$$

The program can be simplified to:

$$e_1 - \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) = 0, \quad \text{(B.19)}$$

$$\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))))$$
$$-(1-e_1) \cdot \delta \left(-e_2 + \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) \right) = 0, \quad \text{(B.20)}$$

$$\delta \cdot (1-e_2) \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))))$$
$$-(1-e_1) \cdot \delta \cdot \left(-(1-e_2) - \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) \right) = 0. \quad \text{(B.21)}$$

Summing (B.20) and (B.21) yields:

$$\begin{aligned}\delta \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) \\ -(w_N + e_2 \cdot (w_L - w_N)) \ + (1 - e_1) \cdot \delta \ &= \ 0. \end{aligned} \tag{B.22}$$

Using (B.19):

$$\delta \cdot e_1 + \delta \cdot (1 - e_1) \ = \ 0, \tag{B.23}$$

which can only be satisfied for the (uninteresting) case that $\delta = 0$. We can thus rule out Case 1.

Case 2 Suppose the solution satisfies the conditions for Case 2. Then it needs to satisfy the following Karush-Kuhn-Tucker optimality conditions:

$$e_1 - \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) - \mu_2 \ = \ 0, \tag{B.24}$$

$$\begin{aligned}\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ -\delta \cdot (1 - e_1) \cdot \left(-e_2 + \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N))\right) + \delta \cdot \mu_2 \ &= \ 0, \end{aligned} \tag{B.25}$$

$$\begin{aligned}\delta \cdot (1 - e_2) \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (e_2 \cdot Y_L - Y_N - (w_N + e_2 \cdot (w_L - w_N)))) \\ -(1 - e_1) \cdot \delta \left(-(1 - e_2) - \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N))\right) \ &= \ 0. \end{aligned} \tag{B.26}$$

We multiply (B.24) by δ, sum all three equations and once more obtain:

$$\delta \cdot e_1 + \delta \cdot (1 - e_1) \ = \ 0, \tag{B.27}$$

which means we can rule out Case 2.

Case 3 Suppose the solution satisfies the conditions for Case 3. This is the case when the participation constraint binds: $-e_1 \cdot w_E + C(e_1) - \delta \cdot (1 - e_1) \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2)) = 0$, and the truth-telling constraint is slack.

Then, it needs to satisfy the following Karush-Kuhn-Tucker optimality conditions:

$$\begin{aligned} & e_1 - \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & + \mu_1 \cdot \left(-e_1 + \frac{\partial e_1}{\partial w_E} \cdot \underbrace{(-w_E + C'(e_1) + \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2)))}_{=0} \right) \quad = \quad 0, \end{aligned} \tag{B.28}$$

$$\begin{aligned} & \delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & \qquad - (1 - e_1) \cdot \delta \cdot \left(-e_2 + \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) \right) \\ & \qquad + \mu_1 \cdot (-(1 - e_1) \cdot \delta \cdot e_2) \\ & + \underbrace{\mu_1 \cdot \left(-\delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} \cdot (-w_E + C'(e_1) + \delta \cdot (w_N + e_2 \cdot (w_L - w_N) - C(e_2))) \right)}_{=0} \\ & \qquad + \underbrace{\mu_1 \cdot \left(\frac{\partial e_2}{\partial w_L} \cdot (-(1 - e_1) \cdot \delta) \cdot \underbrace{(w_L - w_N - C'(e_2))}_{=0} \right)}_{=0} \quad = \quad 0, \end{aligned} \tag{B.29}$$

$$\begin{aligned} & \delta \cdot (1 - e_2) \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & \qquad - (1 - e_1) \cdot \delta \cdot \left(-(1 - e_2) - \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) \right) \\ & \qquad + \mu_1 \cdot (-(1 - e_1) \cdot \delta \cdot (1 - e_2)) \\ & \qquad + \underbrace{\mu_1 \cdot \left(-\delta \cdot (1 - e_2) \cdot \frac{\partial e_1}{\partial w_E} \cdot (\dots) \right)}_{=0} \\ & \qquad + \underbrace{\mu_1 \cdot \left(\frac{\partial e_2}{\partial w_N} \cdot (-(1 - e_1) \cdot \delta) \cdot \underbrace{(w_L - w_N - C'(e_2))}_{=0} \right)}_{=0} \quad = \quad 0. \end{aligned} \tag{B.30}$$

The program can be simplified to:

$$\begin{aligned} & e_1 - \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & \qquad - e_1 \cdot \mu_1 = 0, \end{aligned} \tag{B.31}$$

$$\begin{aligned} & \delta \cdot e_2 \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & \qquad - (1 - e_1) \cdot \delta \cdot \left(-e_2 + \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) + e_2 \cdot \mu_1 \right) = 0, \end{aligned} \tag{B.32}$$

$$\begin{aligned} & \delta \cdot (1 - e_2) \cdot \frac{\partial e_1}{\partial w_E} \cdot (Y_E - w_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N)))) \\ & - (1 - e_1) \cdot \delta \cdot \left(-(1 - e_2) - \frac{\partial e_2}{\partial w_L} \cdot (Y_L - Y_N - (w_L - w_N)) + (1 - e_2) \cdot \mu_1 \right) = 0. \end{aligned} \tag{B.33}$$

Again, we multiply (B.31) by δ and sum all three equations:

$$\begin{aligned} \delta \cdot e_1 \cdot (1-\mu_1) - \delta \cdot (1-e_1) \cdot (-1+\mu_1) &= 0 \\ \Leftrightarrow \delta \cdot (1-\mu_1) &= 0 \\ \Leftrightarrow \mu_1 &= 1. \end{aligned} \tag{B.34}$$

Plugging $\mu_1 = 1$ into (B.31) yields:

$$w_E = Y_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))), \tag{B.35}$$

which can be used to derive from (B.32) or (B.33):

$$w_L - w_N = Y_L - Y_N, \tag{B.36}$$

such that:

$$w_E = Y_E + \delta \cdot w_N. \tag{B.37}$$

The truth-telling constraint requires that:

$$\begin{aligned} & w_E \geq \delta \cdot w_L \\ \Leftrightarrow \quad & Y_E - \delta \cdot (Y_N - w_N) \geq \delta \cdot (Y_L - Y_N) + \delta \cdot w_N, \end{aligned} \tag{B.38}$$

which is violated in the parameter constellation under consideration.

Case 4 We could rule out all cases but Case 4, where the participation constraint and the truth-telling constraints are binding. When the truth-telling constraint binds, $w_E = \delta \cdot w_L$, we effectively have only two variables to choose and we write the objective function to be minimized as:

$$-e_1 \cdot (Y_E - \delta \cdot w_L) - (1-e_1) \cdot \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - (w_N + e_2 \cdot (w_L - w_N))). \tag{B.39}$$

We can write the (binding) participation constraint as:

$$-e_1 \cdot \delta \cdot (w_L - w_N) + C(e_1) - \delta \cdot w_N - (1-e_1) \cdot \delta \cdot (e_2 \cdot (w_L - w_N) - C(e_2)) = 0. \tag{B.40}$$

The binding participation constraint allows us to restate (B.39) as follows:

$$-e_1 \cdot Y_E - (1-e_1) \cdot \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N)) + C(e_1) + (1-e_1) \cdot \delta \cdot C(e_2). \quad \text{(B.41)}$$

We denote $\Delta = w_L - w_N$. The agent's optimization problem can be expressed as follows:

$$e_1 \cdot \delta \cdot \Delta - C(e_1) + (1-e_1) \cdot \delta \cdot (e_2 \cdot \Delta - C(e_2)) + \delta \cdot w_N. \quad \text{(B.42)}$$

such that the optimal effort levels are defined by:

$$\begin{aligned} e_2 &= C'^{-1}(\Delta), && \text{(B.43)} \\ e_1 &= C'^{-1}(\delta \cdot (\Delta \cdot (1-e_2) + C(e_2))). && \text{(B.44)} \end{aligned}$$

We have $e_1 \leq e_2$ if and only if:

$$\begin{aligned} \Delta &\geq \delta \cdot (\Delta \cdot (1-e_2) + C(e_2)) && \text{(B.45)} \\ \Leftrightarrow \delta &\leq \frac{\Delta}{\Delta \cdot (1-e_2) + C(e_2)}. && \text{(B.46)} \end{aligned}$$

As $\delta \leq 1$, the condition above has no bite if

$$\begin{aligned} \Delta &\geq \Delta \cdot (1-e_2) + C(e_2) && \text{(B.47)} \\ \Leftrightarrow \Delta &\geq \frac{C(e_2)}{e_2} && \text{(B.48)} \\ \Leftrightarrow C'(e_2) &\geq \frac{C(e_2)}{e_2}. && \text{(B.49)} \end{aligned}$$

The above inequality is strictly satisfied due to the convexity of $C(\cdot)$. We conclude that $e_1 < e_2$ for all $\Delta > 0$.

Moreover,

$$\begin{aligned} \frac{\partial e_2}{\partial \Delta} &= \frac{1}{C''(\Delta)} > 0, && \text{(B.50)} \\ \frac{\partial e_1}{\partial \Delta} &= \frac{\delta \cdot (1-e_2)}{C''(\delta \cdot (\Delta \cdot (1-e_2) + C(e_2)))} > 0. && \text{(B.51)} \end{aligned}$$

The principal's optimal choice of Δ is characterized by:

$$\frac{\partial e_1}{\partial \Delta} \cdot (-Y_E + C'(e_1) + \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - C(e_2))) + \frac{\partial e_2}{\partial \Delta} \cdot \delta \cdot (1 - e_1) \cdot (-(Y_L - Y_N) + C'(e_2)) = 0. \tag{B.52}$$

The first summand is zero if $C'(e_1) = Y_E - \delta \cdot (Y_N + e_2 \cdot (Y_L - Y_N) - C(e_2))$, which requires $\Delta = \Delta_1^* = \frac{Y_E}{\delta \cdot (1-e_2)} - Y_N - \frac{Y_L - e_2}{1-e_2}$. The second summand is zero if $C'(e_2) = Y_L - Y_N$, which requires $\Delta = \Delta_2^* = Y_L - Y_N$. As she has only one instrument, Δ, at her disposal that affects the agent's effort choices in both periods simultaneously, these effort levels can in general not be implemented. Δ_1^* is smaller than Δ_2^* as $Y_E < \delta \cdot Y_L$.
Using (B.43) and (B.44), (B.52) can be expressed as follows:

$$\frac{\partial e_1}{\partial \Delta} \left(-Y_E + \delta \cdot Y_N + \delta \cdot \Delta + \delta \cdot e_2^S \cdot (Y_L - Y_N - \Delta)\right) + \frac{\partial e_2^S}{\partial \Delta} \cdot \delta \cdot (1 - e_1^S) \cdot (-(Y_L - Y_N) + \Delta) = 0. \tag{B.53}$$

The LHS of (B.53) strictly increases in Δ. Suppose the principal sets $\Delta = \Delta_2^* = Y_L - Y_N$, the spread that induces first best effort in the second period. If $Y_E < \delta \cdot Y_L$, $\Delta^S < Y_L - Y_N$. This implies that $\Delta^S < \Delta_2^* = Y_L - Y_N$ in the optimum, such that $e_2^S < e_2^*$ and $e_1^S > e_1^*$ and deviations from the first-best are inevitable. □

B.2 Multiple Agents - Full Exposition

In the following, we provide a detailed exposition of the results presented in Section 2.4.2. The principal's maximization problem equals:

$$\begin{aligned}(1-(1-e_1^A)\cdot(1-e_1^B))\cdot Y_E - C(e_1^A) - C(e_1^B) \qquad \text{(B.54)}\\ +(1-e_1^A)\cdot(1-e_1^B)\cdot\delta\cdot(Y_N+(1-(1-e_2^A)\cdot(1-e_2^B))\cdot(Y_L-Y_N)-C(e_2^A)-C(e_2^B))\end{aligned}$$

which can be rewritten as (2.17). First, we will show by means of an example that parameter constellations exist that satisfy our assumptions and yield a symmetric solution to the principal's problem. Therefore, we suppose that our cost function is defined by $C(e_t) = {e_t}^2$ and show that the desired properties are fulfilled for this particular case.
Given its structure, it is permissible to treat the problem as two separate bivariate optimization problems with decision variables e_t^A and e_t^B for any t. We start with $t = 2$. From the first-order condition, it follows that the optimal effort levels for agent A and B, respectively, are defined by

$$(1-e_2^B)\cdot(Y_L-Y_N) = 2\cdot e_2^A. \qquad \text{(B.55)}$$

and:

$$(1-e_2^A)\cdot(Y_L-Y_N) = 2\cdot e_2^B. \qquad \text{(B.56)}$$

Because $C'(1) > Y_L - Y_N$, the optimal e_2^A (conditional on e_2^B) is below 1 for $e_2^B = 0$ and equal to 0 for $e_2^B = 1$ and vice versa. Hence, there must be a unique intersection at $e_2^A = e_2^B = \frac{(Y_L-Y_N)}{(Y_L-Y_N+2)}$ which is the unique and symmetric candidate for an optimum.
Given the symmetry of efforts, the problem becomes in fact an univariate optimization problem and the sufficient conditions for an optimum to exist are

$$-C''(e_2^A) < 0 \qquad \text{(B.57)}$$

and

$$C''(e_2^A)^2 - (Y_L-Y_N)^2 > 0. \qquad \text{(B.58)}$$

The fulfillment of both conditions can be readily verified.
For $t = 1$, the rationale for interior, unique and symmetric solutions is exactly the same as for $t = 2$, with the sole exception that "$Y_L - Y_N$" has to be replaced by "$Y_E - \delta\cdot V_2^{**}$", where V_2^{**} refers to the maximized surplus in the second period when employing two

agents.

Like in the case with only a single agent, we assume that $Y_E - \delta \cdot V_2^{**} > 0$ must hold, a condition that is more restrictive than in the single-agent case, because $V_2^{**} > V_2^*$.

Having verified that the case in which the principal wants to employ two agents with identical effort levels is relevant in our setting, it suffices to look at agent A's incentives. When the second period is reached, the agent chooses e_2^A so as to maximize:

$$e_2^A \cdot w_L + (1 - e_2^A) \cdot w_N - C(e_2^A), \tag{B.59}$$

with its solution characterized by:

$$C'(e_2^A) = w_L - w_N. \tag{B.60}$$

Bringing incentives into line to induce the socially desired effort requires:

$$w_L - w_N = (1 - e_2^B) \cdot (Y_L - Y_N). \tag{B.61}$$

Supposing we implement that spread, and making use of the period-wise symmetry of efforts, the agent's expected rent in $t = 2$ (if reached) is:

$$R_2 = w_N + e_2^A \cdot (1 - e_2^A) \cdot (Y_L - Y_N) - C(e_2^A), \tag{B.62}$$

and the maximized surplus in the second period is:

$$V_2^{**} = Y_N + (1 - (1 - e_2^A)^2) \cdot (Y_L - Y_N) - 2 \cdot C(e_2^A), \tag{B.63}$$

so that:

$$\begin{aligned} V_2^{**} - R_2 &= Y_N + (1 - (1 - e_2^A)^2 - e_2^A \cdot (1 - e_2^A)) \cdot (Y_L - Y_N) - C(e_2^A) - w_N \\ &= Y_N + e_2^A \cdot (Y_L - Y_N) - C(e_2^A) - w_N. \end{aligned} \tag{B.64}$$

The agent chooses e_1^A so as to maximize (2.20) with its solution characterized by:

$$C'(e_1^A) = w_E - (e_1^B \cdot w_F + (1 - e_1^B) \cdot \delta \cdot R_2). \tag{B.65}$$

Bringing incentives into line to induce the socially desired effort requires:

$$w_E - (e_1^B \cdot w_F + (1 - e_1^B) \cdot \delta \cdot R_2) = (1 - e_1^B) \cdot (Y_E - \delta \cdot V_2^{**}), \tag{B.66}$$

so that:

$$w_E = e_1^B \cdot w_F + (1 - e_1^B) \cdot (Y_E - \delta \cdot (V_2^{**} - R_2)). \tag{B.67}$$

Truth-telling is ensured by (2.21), that is:

$$\begin{aligned}
& e_1^B \cdot w_F + (1 - e_1^B) \cdot (Y_E - \delta \cdot (V_2^{**} - R_2)) \geq e_1^B \cdot w_F + (1 - e_1^B) \cdot \delta \cdot w_L \\
\Leftrightarrow \quad & Y_E - \delta \cdot (Y_N + e_2^A \cdot (Y_L - Y_N) - C(e_2^A) - w_N) \geq \delta \cdot w_L \\
\Leftrightarrow \quad & Y_E - \delta \cdot (Y_N + e_2^A \cdot (Y_L - Y_N) - C(e_2^A) - w_N) \geq \delta \cdot ((1 - e_2^A) \cdot (Y_L - Y_N) + w_N) \\
\Leftrightarrow \quad & Y_E \geq \delta \cdot (Y_L - C(e_2^A)) \qquad \text{(B.68)}
\end{aligned}$$

where (B.68) is referred to as (2.22) in the main text.

3 The Inspection Game in Science*

Matthias Verbeck†
University of Marburg

Abstract

What are the conditions under which fraudulent or erroneous research arises and survives in the scientific community? To answer this question, we build on the work of Lacetera and Zirulia (2011) and model the scientific approval process along the lines of an inspection game. A researcher publishes a possibly fraudulent or faulty result which comes under scrutiny from a (large) scientific readership. Scrutinizing scientific publications may constitute a public good for the scientific community, such that the volume of (unrevealed) faulty research can *increase* with the number of interested readers. In fact, an author might intentionally increase the level of fraud so as to attract more readers, thereby aggravating the free rider problem and *reducing* the likelihood of getting caught. Moreover, the model sheds light on the question of whether and when a greater diversity of opinions in the scientific community helps to weed out flawed research.

Keywords: Scientific Misconduct, Reproducibility, False Positives, Inspection Game, Volunteer's Dilemma, Economics of Science

JEL codes: A14, D82, K42, O31, Z1

*I wish to thank Elisabeth Schulte, Rebecca Dietrich, the participants in the internal research seminar run by the University of Marburg's Economics department, and the participants in the 2018 MAGKS research seminar in Rauischholzhausen for their very helpful comments and suggestions. I also thank Guido Buenstorf and Stephan B. Bruns for familiarizing me with the work of Lacetera and Zirulia as part of a superb Ph.D course on the "Economics of Science" at the University of Kassel in 2016/17.

†Corresponding author: matthias.verbeck@gmail.com

3.1 Introduction

In the more recent past, several cases of flawed academic publications in highly respected journals have attracted attention not just from the scientific community but even from the public at large. Among the most notable cases was an article of Hwang et al. (2005), published in *Science*, in which the authors claimed to have succeeded in generating human embryonic stem cells through cloning. Several months later, after a couple of researchers had unsuccessfully tried to replicate the results, the article was retracted, and the findings were ultimately exposed as fraudulent. Another widely noted case from the field of economics involved not scientific misconduct, but mere human error. "Growth in a time of debt" (Reinhart and Rogoff 2010), published in *American Economic Review: Papers & Proceedings* analyzed the connection between national debt levels and economic growth rates, and concluded that "for levels of external debt in excess of 90 percent" GDP growth was "roughly cut in half". Their article was widely cited and provided a rationale for austerity measures in debt-ridden economies. In 2013, however, graduate student Thomas Herndon discovered that the reported size of the effect was highly exaggerated for the trivial reason that an Excel sheet was flawed (Herndon et al. 2014).

Another important reason why scientific publications might contain defective results is not fraud or error, but rather the scientific journal's prevailing selection process, which favors the publication of statistically significant results. In his seminal paper on reporting bias, Ioannidis (2005) argued that the (vast) majority of claimed research findings are false, at least in the field of medical or medical-related research. Recent studies (e.g. Baker 2016) confirm the existence of a "replication crisis" which indicates that published results are less deserving of trust, unfortunately, than scientists would like. In summary, then, it can be said that a substantial number of papers manage to clear the hurdle of peer review and get published even though they contain false, fraudulent or at least non-reproducible findings.

The crucial question, then, is whether academia succeeds in weeding out such false results over time. In other words: Do the wrong results that made it into academic journals also finally find their way into academic textbooks? In fact, academia is a realm characterized by a high degree of autonomy and a rather low level of external intervention. The role of the individual researcher is therefore complex, as s/he is a contributor, competitor and

supervisor, all at the same time. Despite this complexity, the individual aspiration to gain a reputation and the resulting competitive pressure among peers is often considered sufficient to eliminate wrong or deficient findings over time (e.g. Merton (1973)). The model presented here aims to aims to add clarity to the question of whether and when this notion is justified.

One of the first theoretical analyses of scientific misconduct is the enlightening model of Lacetera and Zirulia (2011) (henceforth "L&Z"). Their findings include, but are not limited to, the following:

- Cases of *detected* fraudulent research are not representative of the *overall* amount of bogus research, since less innovative research papers are not likely to be scrutinized at all.
- A reduction in the individual cost of checking scientific results does not necessarily lead to an increase in detected fraudulent research.
- High pressure to publish meaningful results may decrease (and not increase) scientific misconduct, since peers will then check results with increased probability.

While their analysis is extremely insightful, it leaves the (crucial) role of audience size and structure unmodeled. In their model, the scientific audience is assumed to consist of a single reader who may or may not check a published article for soundness. This is of course an extreme simplification since a typical scientific publication will attract a wider readership, especially so when the published result is of greater importance. Without thorough analysis, it remains unclear whether and how the existence of multiple readers affects the volume of (undetected) flawed findings. A cursory view inspired by the economic theory of crime (Becker 1968) suggests that larger audiences will unambigiously help to keep science clean. If n readers check a given article and inform the scientific community of any flaw or fraud they find - if present - with fixed individual probability of $\tau \in (0,1)$, then the overall probability of detection is $1-(1-\tau)^n$ and therefore strictly increases in n. This result would be a desirable one since it would mean that highly influential articles (those with many readers) survive in the scientific community if, and only if, the published findings are valid. Furthermore, if it were true that the overall probability of fraud detection increases with audience size, then a higher number

of readers should be effective in preventing an author from committing fraud in the first place.

However, as the critique of Tsebelis (1989, 1990) has shown, a norm-enforcing authority should be modeled not as a fixed probability distribution, but rather as a rational player. If scrutinizing an article for flaws or fraud involves a private cost for the reader, then - given the presence of a multitude of readers - this activity constitutes a public good for the scientific community, and incentives to free ride on the efforts of peers must be taken into consideration. In fact, L&Z also speculate that free riding could affect individual behavior when a multitude of readers is assumed (p. 594). Our analysis explicitly addresses this issue and can therefore be regarded as the conflation of a public good game and an inspection game.

Moreover, we use their framework to also analyze the prevalence of erroneous or non-reproducible (but not fraudulent) findings in academia. The crucial action on the part of the researcher is the amount of effort they exert. A scientist who has invested a fair amount of time double-checking her results is less likely to unwittingly submit a flawed paper than her less diligent colleague. Or else, a researcher who makes a greater effort in data collection is less prone to spuriously produce a false positive result than her colleague who uses a smaller sample size.

The model presented here therefore allows us to analyze both fraudulent and flawed research in a common setting that sheds light on the prevalence of problematic research findings in academia. It is mainly designed to capture aspects of empirical academic research, but in principle the model can also be applied to any kind of science where errors or deception can occur, e.g. mathematics or theoretical physics.

The remainder of the paper is structured as follows: Section 3.2 provides an overview of the literature related to our research. Section 3.3 presents the basic model. The publication process is modeled as an extensive-form game under incomplete information where we first analyze the case of fraudulent research (the "deception game") before we treat the problem of erroneous research (the "delusion game"). Section 3.4 extends the derived results in numerous ways. Most importantly, we analyze how players' behavior is affected by competition among readers, the possibility of multiple audiences, heterogeneity among readers, and the existence of an editor who might also wish to check articles. Section 3.5 discusses our findings, and section 3.6 concludes. Detailed proofs can be found in

Appendix C.

3.2 Related Literature

As stated above, our work is mainly an extension of the model of Lacetera and Zirulia (2011). We make extensive use of their framework and also adopt most of their notation. The crucial difference between their model and ours is that we do not limit the number of readers to one and explicitly address the positive externality that any scrutinizing reader creates for the entire scientific community. Moreover, while their model mainly focuses on different types of research and their respective vulnerability to (undetected) fraud, we are interested in how the size and structure of the academic readership influence the volume and persistence of fraudulent publications. Our work is also closely related to a follow-up paper by Kiri, Lacetera and Zirulia (2018). In this work, the authors explicitly model a researcher's effort decision if a colleague wishes to scrutinize a publication. In an extension of their model, they increase the number of peers to two and show that this increase can reduce an author's incentive to strive for high-quality research. Moreover, the overall probability of detecting deficient findings decreases. Even if their findings might seem similar to ours at first glance, the underlying mechanism is different, and free riding is not considered in either of the two papers. As we proceed, we will repeatedly highlight the differences between our model and theirs and discuss them again in greater detail in section 3.5.

Altogether, our paper is related to the theoretical literature on questionable or fraudulent research practices as well as the (problems inherent in the) academic publication process. A still very readable overview of different forms and shades of academic misdemeanor is offered in LaFollette (1992). Wible (1998) provides a first formal analysis of the academic publication process that is situated in decision theory rather than game theory. Among others, Ioannidis (2005, 2012) and Bettis (2012) argue very forcefully that a vast fraction of published research articles will contain false positive results which might remain unchallenged. Bobtcheff et al. (2017) present a formal analysis of academic publishing and show how the researchers' striving for priority can undermine their incentives to be concerned about quality. McElreath and Smaldino (2015) and Nissen et al. (2016) model the academic approval process and analyze conditions under which incorrect claims will falsely be adopted by the scientific community. Gall and Maniadis (2019) provide a model

in which competing authors can choose between different levels of transgression, such as omission of data as opposed to overt data fabrication. They find that policies that aim to prevent mild forms of misdemeanor are also suited to prevent more severe forms of scientific misconduct, but not vice versa. Furthermore, our work is related to the class of "persuasion games" that analyze the difficulties in honestly transferring scientific findings between asymmetrically informed parties (e.g. Felgenhauer and Schulte 2014, Henry and Ottaviani 2017, Di Tillio et al. 2017).

Moreover, our model features characteristics of a typical "inspection game" (Tsebelis 1989, Andreozzi 2004). In its most simple version, a potential wrongdoer can either act in a way preferred by the inspector (e.g. working hard) or contrary to the inspector's wishes (e.g. shirking hard work). The inspector, for his part, either does or does not engage in costly monitoring and will discover misbehavior only if he decides to monitor. Typically, this game has only one equilibrium in mixed strategies, in which the inspector only sometimes checks the agent, who in turn cheats with positive probability. Interestingly, a higher level of punishment for the perpetrator does not affect the likelihood of cheating, but instead reduces the inspector's incentives to engage in monitoring. In our model, each reader represents a potential "inspector" who can check the soundness of a colleague's work. Scrutinizing an author's work then constitutes a public good for the scientific community and provokes a "volunteer's dilemma" (Diekmann 1985) among all colleagues. Such a dilemma is characterized by the fact that the provision of a (public) good only takes place with certainty as long as the number of potential contributors is restricted to one. Once there is a multiplicity of potential contributors, every player provides the public good with a probability strictly smaller than 1 in the symmetric equilibrium. Hence, a diffusion of responsibility takes place, and the provision of the public good could fail altogether. We are explicitly interested in the extent to which this free rider problem affects a rational author's incentives to cheat in the first place.

Empirical works that deal with unreplicable, questionable or fraudulent research are abundant, and the contributions cited here are only exemplary. The problem of non-reproducible research is especially well documented in medicine (e.g. Begley and Lee (2012)) and psychology (e.g. Simmons et al. (2011) and Wagenmakers et al. (2011)), but recent studies also call into question the replicability of other disciplines, such as (experimental) economics (Camerer et al. (2016) and Brodeur et al. (2016)) and management

science (Goldfarb and King (2016)). Bruns et al. (2019) find evidence of errors and biases in reported significance levels in innovation research. Necker (2014) provides a survey conducted among members of the European Economic Association, a non-negligible fraction of whom admitted that they had already engaged in questionable research practices. Furman and co-authors (2012) show that after an academic publication is retracted, it will be cited far less often in the future, therefore supporting the idea that word spreads fast in the scientific community. Azoulay et al. (2015) show that fraudulent articles may contaminate whole fields of research and are therefore suited to shift future research activities.

3.3 The Model

3.3.1 Fraudulent Research: The Deception Game

In the baseline model, there are two types of player: a (male) author (A) who produces a scientific article and a (possibly large) readership[15], consisting of $n \geq 2$ (female) readers (R) who can scrutinize a published article. In the extensions section, we will furthermore introduce a (male) editor (E) who can also check publications for soundness and who is interested in publishing only solid results.

The game consists of three stages. At stage 1, nature determines the researcher's output level. The researcher's output Y is modeled as binary, where a success is labeled as S, and a failure is denoted as F. The probability of success is denoted by β, whereas $1 - \beta$ defines the probability of failure. Although not explicitly modeled here, we can understand this probability of succcess as the result of some positive effort level that the researcher found optimal to invest at some earlier (unmodeled) stage. The probability of success is common knowledge to all players. The *realized* output level, however, is A's private information and cannot be directly observed by any of the other players. At stage 2, A decides whether and how to present the research output to the scientific community. The set of actions depends on the observed output level Y. For output level S, A can decide to publish (*pub S*) or not to publish (*no pub S*) the article. For output level F, however, publication of the resulting article - in its current form - is not possible.[16]

[15] We will use the terms audiences and readerships interchangeably.

[16] This assumption is motivated by the fact that null results are much more difficult to publish, especially in highly ranked scientific journals.

Hence, A might decide to embellish the results. This means that the true type of output Y and the announced type of output $\widehat{Y}$ may differ. He therefore chooses between $(pub\ \widetilde{S})$ which he plays with probability p, and $(no\ pub\ F)$, played with probability $(1-p)$. The former action refers to the practice of "overselling" a result, i.e. making it seem a success, even though it is not one. Naturally, we assume that this behavior is at odds with the scientific community's code of conduct. We can think of actions like deliberately applying inappropriate statistical methods, the unjust deletion of outliers, outright fabrications of data or any other wrongful behavior that yields a higher level of statistical significance or bestows the work with an undue amount of scientific recognition. If A decides to publish the fraudulent article, the game enters stage 3. At this stage, every reader i simultaneously either chooses to check (*check*) the article with individual probability q_i, or not to check (*no check*) the article with probability $1-q_i$. If a reader decides to check an article, she will detect scientific misconduct with probability $\tau > 0$.[17] We assume that as soon as at least one researcher detects fraud, word will spread within the scientific community and the fact will become common knowledge. This assumption seems justified when the checking reader informs the journal editor, who would then usually retract the article. Or the reader writes an article of her own exposing the author's misdemeanor. Like in L&Z, the *check*-action describes different behaviors, e.g. spot-checking statistical figures for obvious inconsistencies, but also replicating the results with the author's original data, or conducting an experiment similar to the author's, etc.

Next, we describe the payoff structure. If A publishes a non-fraudulent article, he will gain a benefit $B > 0$. This benefit represents gains in reputation, advanced career opportunities and the like. The fact that B is positive implies that the author will always publish a solid article if $Y = S$. If the author instead publishes a faked article that remains unchallenged, he will gain a benefit B', and we assume that $B \geq B' > 0$. Should the author decide to publish no article at all, he will receive a zero payoff. The assumption that $B' > 0$ implies that the author's motivation to publish articles is mainly driven by career concerns rather than by promoting the state of the art.

Publication of a research article also generates a payoff for every reader. In contrast to L&Z, we assume that this payment is state-dependent and let W denote any reader's payoff in case the author's contribution is valid, whereas W' denotes the payoff when the

[17] This assumption differs from L&Z, who assume that a check is certain to uncover fraudulent behavior.

published results are fake. We assume that $W \geq W'$. Moreover, W can take any value in $\mathbb{R}$, whereas $W' < 0$.[18] Our assumptions that $W' < 0$ and $W' < W$ can be motivated by the fact that most (honest) researchers - like people in general - do not like being cheated and clearly prefer no publication or a sound publication over a false publication. In addition, the rejection of faked results can also be justified by the reader's position as a producer of new research. Fraudulent results can certainly mislead a reader to pursue wrong or unpromising paths in her future research and therefore have a negative effect on a reader's utility.

If any R decides to check a published result, she has to bear cost $k > 0$.[19] If a reader checks an article and finds scientific misconduct, she will gain a benefit $E(G) > 0$. In general, this benefit is likely to depend on the number of other researchers who also managed to successfully discover the fraud. A single reader's gain (in reputation) might be smaller if more colleagues successfully uncover a wrongdoer. This assumption can be justified by the priority principle in science. Only if a reader manages to be the first to prove the invalidity of a previously accepted result will she gain in reputation; otherwise, she will usually come away empty-handed. We explicitly address this issue in section 3.4.1. In the following baseline model, however, we make the simplifying assumption that $E(G) = G$, meaning that it is independent of the (expected) number of successful readers. Completing the list of payoff parameters is the cost that the author experiences if he is caught cheating (i.e. loss of reputation, monetary fines, etc.). We denote this cost by $g > 0$. Should A prefer to cheat when his research project remaines fruitless, his expected payoff at $t = 0$ equals

$$\beta \cdot B + (1-\beta) \cdot \left(\left(\prod_{i=1}^{n} (1 - q_i \cdot \tau) \right) \cdot B' + \left(1 - \left(\prod_{i=1}^{n} (1 - q_i \cdot \tau) \right) \right) \cdot (-g) \right). \quad (3.1)$$

[18] Whether W is positive or not generally hinges on the readers' perception of the published result. If a reader perceives the work to be complementary to her own prior work, or if the published result simply supports a view which the reader approves, W will be positive. If the contribution is regarded as a substitute for a reader's prior or future research, however, W will be negative. Hence, a publication that yields negative values for W can be considered to be at odds with the audience's preferences and to contradict existing theories. High values for W, on the other hand, represent results that fit in well with the existing research, do not limit the scope for readers' for own contributions, and teach the reader something that could be of interest to her own research.

[19] This cost can be understood as the obvious cost of collecting data or conducting an experiment, but it can also involve the reader's opportunity cost of not doing original research instead of reviewing pre-existing research.

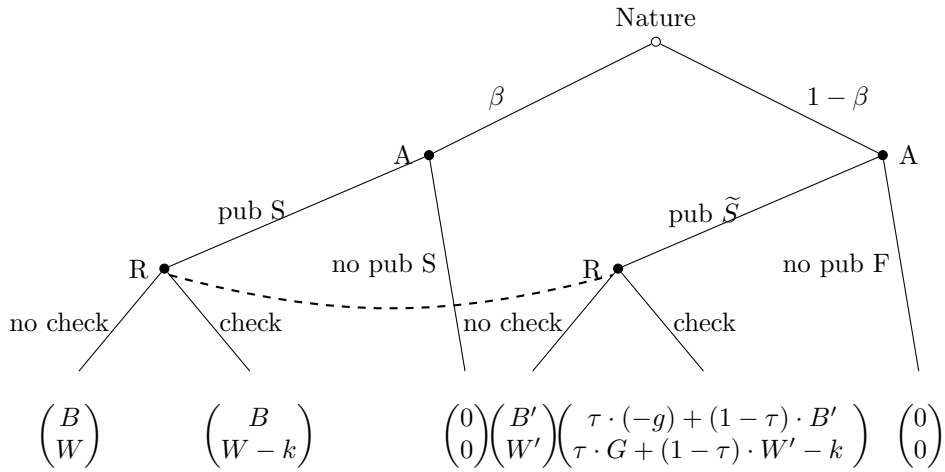

Figure 3.1: The deception game under the simplified assumption of $n = 1$ reader

Likewise, the total expected utility of any reader willing to check a publication is

$$\mu \cdot W + (1-\mu) \cdot \left(\left(\prod_{i=1}^{n} (1 - q_i \cdot \tau) \right) \cdot W' + \tau \cdot G \right) - k \tag{3.2}$$

and μ denotes the readers' updated probability estimation on an author's success, after an article has been published.

Unlike in the contributions of L&Z and Kiri et al. (2018), a checking reader creates a positive externality for the whole scientific community. With each additional scrutinizing colleague, the expected gain of any reader who checks a publication herself will decrease. We add one final assumption that restricts our analysis to cases where this externality is so large that *not all* readers will strictly prefer to check. This is guaranteed if

$$\begin{aligned} \beta \cdot W + (1-\beta) \cdot (1-\tau)^{n-1} \cdot W' > \beta \cdot W + (1-\beta) \cdot (\tau \cdot G + (1-\tau)^n \cdot W') - k \\ \Leftrightarrow G < (1-\tau)^{n-1} \cdot W' + \frac{k}{(1-\beta) \cdot \tau}. \end{aligned} \tag{3.3}$$

Figure 3.1 shows the game tree under the simplifying assumption that there is only one reader.

The appropriate solution concept for the presented game is that of a perfect Bayesian equilibrium. We solve for all *symmetric* perfect Bayesian equilibria and distinguish between equilibria in pure and mixed strategies.

An equilibrium is fully characterized by (a) the author's publication decision in case of

success (*pub S* or *no pub S*), (b) the author's publication decision in case of failure (*pub* or *no pub F*), (c) the action chosen by any reader in case an article gets published (*check* or *no check*), and (d) the readers' posterior belief μ about the author's success in case of publication.

Proposition 3.1. *For the deception game, every parameter constellation yields exactly one symmetric equilibrium, such that*

1. *For* $G < W' + \frac{k}{(1-\beta)\cdot\tau}$*:* $p = 1$, $q_i = 0$, $\mu = \beta$.

   The probability that a published article is fraudulent is $(1-\beta)$ *and the probability that a fraudulent article gets caught (if published) is 0. We call this equilibrium "Pooling I".*

2. *For* $G \geq W' + \frac{k}{(1-\beta)\cdot\tau}$ *and* $(1-\tau)^n > \frac{g}{B'+g}$*:* $p = 1$, $q_i \in (0,1)$, $\mu = \beta$.

   Specifically, we have $q_i = \left(1 - \sqrt[n-1]{\frac{1}{W'} \cdot \left(G - \frac{k}{(1-\beta)\cdot\tau}\right)}\right) \cdot \frac{1}{\tau}$*. The probability that a published article is fraudulent is* $(1-\beta)$ *and the probability that a fraudulent article gets caught (if published) is* $1-(1-q_i\cdot\tau)^n = 1 - \left(\frac{1}{W'} \cdot \left(G - \frac{k}{(1-\beta)\cdot\tau}\right)\right)^{\frac{n}{n-1}}$*. We call this equilibrium "Pooling II".*

3. *For* $G \geq W' + \frac{k}{(1-\beta)\cdot\tau}$ *and* $(1-\tau)^n \leq \frac{g}{B'+g}$*:* $p \in [0,1]$, $q_i \in [0,1]$, $\mu = \frac{\beta}{\beta + p\cdot(1-\beta)}$.

   Specifically, we have $p = \frac{\beta}{1-\beta} \cdot \frac{k}{\tau\cdot\left(G - W'\cdot\left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}}\right) - k}$ *and* $q_i = \left(1 - \sqrt[n]{\frac{g}{B'+g}}\right) \cdot \frac{1}{\tau}$*. The probability that a published article is fraudulent is* $(1-\beta)\cdot p$ *and the probability that a fraudulent article gets caught (if published) is* $1-(1-q_i\cdot\tau)^n = 1 - \frac{g}{B'+g}$*. We call this equilibrium "Semi-Separation".*

In case of success, the author will publish an article in any of the equilibria.

Proof: See Appendix C.

Similar to L&Z, in equilibrium, fraud will occur with positive probability. The existing equilibria can be characterized along different parameter thresholds. If $G < W' + \frac{k}{(1-\beta)\cdot\tau}$,

not a single reader will want to check, as the expected benefit from doing so would not cover the cost. Since all readers will abstain from checking a publication, a rational author will never stop short of scientific misconduct as the probability of being debunked equals zero.

If $G \geq W' + \frac{k}{(1-\beta)\cdot\tau}$, the readers will check a published article with positive probability.[20] Then, the size of B' relative to g and τ will determine the author's strategy. For $(1-\tau)^n > \frac{g}{B'+g}$, the author's expected punishment is not sufficiently severe to deter him from releasing a fraudulent article, and we observe pooling behavior once more ("Pooling II"). For the readers, the game then essentially turns into a public good game, and the volume of scrutiny does not depend on the author's payoff parameters.[21] Lower values for G and τ and higher values for W', β and k reduce the individual probability of a reader checking a published result. Hence, if most publications can generally be trusted, checking costs are high, and the readers' gains from refuting the article are limited, there might be a good chance that a cheating author will be able to escape undetected. When G increases (assuming that all other variables are held constant), all readers will eventually want to check a publication with probability 1 - that is, when $G \geq W' \cdot (1-\tau)^{n-1} + \frac{k}{(1-\beta)\cdot\tau}$, a parameter constellation we ruled out by assumption. Note that for parameters set between these two extremes (all readers want to check, or no reader wants to check), there are many more asymmetric equilibria where readers differ in their individual probability of examining a finding.

For $(1-\tau)^n \leq \frac{g}{B'+g}$, the author is kept indifferent between cheating and not cheating, and both actions occur with positive probability. Hence, we have a semi-separating equilibrium, and the probability that a published article is fraudulent is lower than in any of the pooling equilibria. Unlike in the second pooling equilibrium ("Pooling II"), positive probabilities for playing *check* now let the author abstain from cheating with positive probability. Therefore, this equilibrium rather resembles the canonical inspection game's equilibrium, in which cheating is not a dominant strategy for the (potential) perpetrator. In particular, the author chooses to cheat more often as k, W' and β increase and as G and τ decrease. The readers only respond to the author's payoff variables and check a publication more frequently as B' increases and g decreases. Note that also in the case

[20] The condition makes checking for at least one reader profitable.

[21] To be more precise, the game gets the structure of a volunteer's dilemma (Diekmann 1985), where the public good would be provided with certainty if there was only one potential contributor.

of $p \in (0, 1)$, there are other asymmetric equilibria where readers check publications with dissimilar probabilities. Figure 3.2 illustrates the three different equilibria for varying values of G and B'.

For given values of B, B' and g, the author's expected payoff is highest if a pooling equilibrium of type one ("Pooling I") emerges, second highest if a pooling equilibrium of type two ("Pooling") emerges, and lowest in the semi-separating equilibrium. This ordering directly corresponds to the respective probability of being caught, which is highest in the semi-separating equilibrium and 0 in the pooling equilibrium of type one.

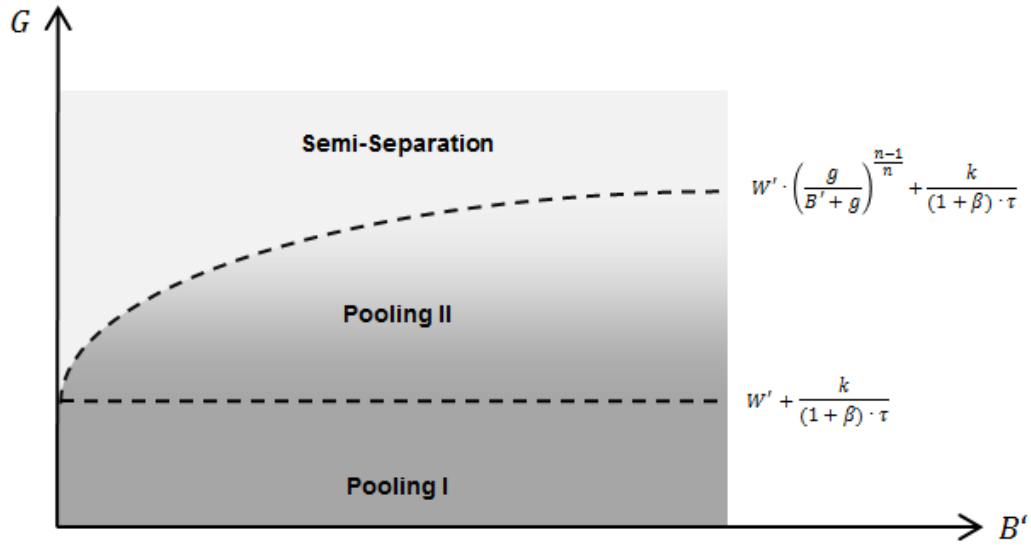

Figure 3.2: Resulting equilibria for different realizations of G and B'. Brighter shades refer to a higher overall probability of fraud detection, given that a publication has been released.

We are mostly interested in how audience size affects the share of debunked fraudulent research, as well as the total volume of fraudulent research. We obtain the counterintuitive result that the volume of both fraud in general and undetected fraud will increase weakly with the number of readers. This holds true for both kinds of equilibria, the two pooling equilibria and the semi-separating equilibrium.

Proposition 3.2. *In the deception game, considering symmetric equilibria, an increase in the number of readers from n to $n+1$ affects the equilibria as follows:*

1. *For* $G < W' + \frac{k}{(1-\beta)\cdot\tau}$*: The absolute volume of fraud remains at* $(1-\beta)$ *and the volume of undetected fraud remains at 0.*

2. *For* $G \geq W' + \frac{k}{(1-\beta)\cdot\tau}$ *and* $(1-\tau)^n < \frac{g}{B'+g}$*: The absolute volume of fraud remains at* $(1-\beta)$ *and the level of undetected fraud increases.*

3. *For* $G \geq W' + \frac{k}{(1-\beta)\cdot\tau}$ *and* $(1-\tau)^n \geq \frac{g}{B'+g}$*: Both the absolute volume of fraud and the volume of undetected fraud increase.*

4. *The parameter set for which "Pooling II" exists increases, and the parameter set for which "Semi-Separation" exists decreases.*

Proof: See Appendix C.

Despite the fact that we have a larger supply of readers and therefore potentially a higher level of scrutiny, the de facto level of checking decreases due to free riding behavior. This implies that the overall probability of fraud detection will never exceed τ, no matter how large the readership is. As n grows in size, the volume of misconduct remains unchanged or even increases.[22] These results contradict common sense beliefs about the academic publication process and show that the notion of a self-correcting scientific community may not be justified. However, caution should be exercised. A larger audience size might also imply different values for all other (payoff) parameters (see also section 4.3 in Kiri et al. (2018)). In particular, it is reasonable to assume that B' and G are higher for larger audiences, thus encouraging the readers' scrutiny and deterring fraudulent behavior. Therefore, we can only conclude that a large number of readers *alone* is not a sufficient condition for a high level of quality in scientific publications.

With a higher level of n, we also observe a shift in the occurring equilibria. The set of parameter values causing the "Pooling II"equilibrium to emerge grows at the expense of the set of parameter values that imply the existence of the semi-separating equilibrium. The validity of the first pooling equilibrium is not affected by a higher n.

What may come as a surprise is that τ does not affect the overall probability of detecting flawed articles *within* the semi-separating equilibrium. A lower τ is always compensated for by a higher q_i, leaving the overall detection probabilities unaffected. Instead, a higher

[22]The increase will only occur in the semi-separating equilibrium. The author will publish fraudulent articles more often, to make the additional reader indifferent between checking and not checking as well.

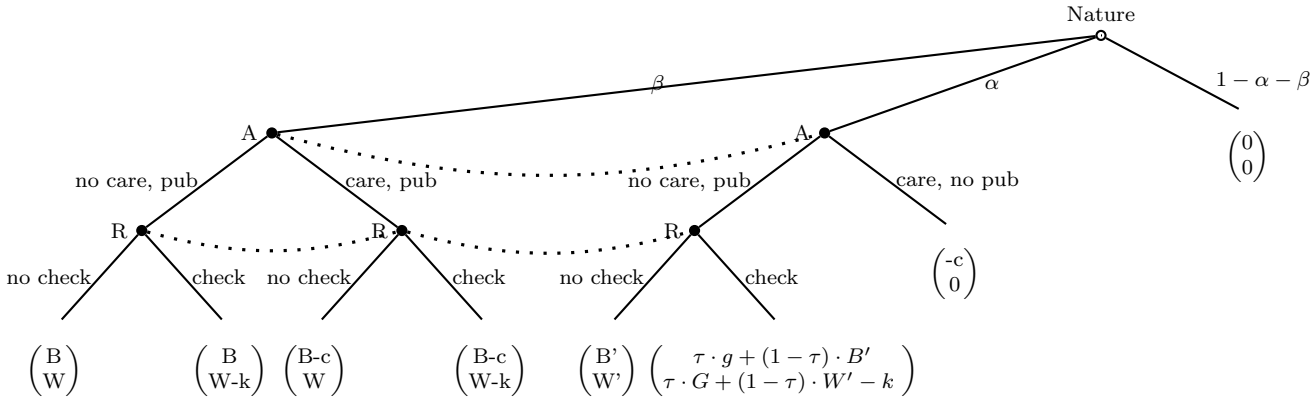

Figure 3.3: The delusion game under the simplified assumption of $n = 1$ reader

value of τ leads to a downward shift in the threshold functions, as depicted in Figure 3.2. As a consequence, the set of parameter values that result in the first pooling equilibrium shrinks, whereas the set of parameter values that imply the semi-separating equilibrium grows. It is furthermore interesting to see that W (in contrast to W') has no influence on the players' behavior.

3.3.2 Erroneous Research: The Delusion Game

In this section, we disregard the possibility of deceiving the scientific community. Instead, we make the assumption that the author is honest, but might unwittingly produce a result that is not replicable or at least less potent than originally claimed. The reason for this can either be the author's individual negligence, or else the result is simply a false positive one. We therefore adjust the presented game in the following way. At stage 1, nature determines the author's *observed* output level as well as the *actual* output level. With probability β, the author experiences and observes a true success. With probability α, the author erroneously observes a success when in fact a failure has been produced, henceforth referred to as "spurious success". Then, with counter probability $1 - \beta - \alpha$, the author rightly observes a failure.[23] In this case, the game ends at stage 1 and all players will receive a zero payoff. For the readers, neither the observed nor the actual output level is known.

Should the author observe a success (true or spurious), the game enters stage 2. Here, the author chooses his binary level of care (*no care* or *care*), and p now denotes the probability

[23]For simplicity, we rule out type II errors, where the author observes a failure even though the result is actually a success.

of not applying care. Investing care means anything that helps ruling out non-replicable or oversold results. For example, to avoid individual errors, the author could consult colleagues to clarify whether a statistical method has been applied correctly, or he double-checks all the data processed by his student assistants. Or else, the author increases the sample size to improve the robustness of his findings. Investing care comes at a cost of $c > 0$. Should the author decide to invest care, he will identify a spurious success with certainty.[24] Then, a publication will only be released in case of a true success, otherwise the author dispenses with making a publication while all readers once more obtain a zero payoff. Should the author decide to not invest care, the article is published in any case, and the author takes the risk of accidently having released a faulty article.

If an article is published, the game enters stage 3, in which the audience can again decide to check or not check the result and will find errors with individual probability τ. Note again that the reader can neither observe the actual and the author's observed output level nor the decision as to whether the author has invested care into his publication.

The players' payoffs are determined analogously to the deception game, though other values for these parameters now seem reasonable (for example, it is plausible that g, the author's utility loss in case a wrong result is detected, is generally milder). Most importantly, we now assume that $B' < 0$ and $B' > -g$, i.e. an author who knows about the article's flaws would never wish to publish it, and should the flaws be detected, his utility loss would be larger than if his error remained unseen by the readers. Moreover, like in the deception game, we assume that not all readers strictly prefer to check. The game tree depicted in Figure 3.3 illustrates the course of action, again under the simplified assumption of a single reader.

The game's equilibria are characterized by (a) the action chosen by the author (*care* or *no care*), (b) the action chosen by any reader in case an article is published (*check* or *no check*), (c) the author's posterior belief μ_A about the research outcome after a success has been observed, and (d) the readers' posterior belief μ_R about the publication's soundness if a publication is released.

Proposition 3.3. *For the delusion game, every parameter constellation yields exactly one symmetric equilibrium, such that*

[24] Qualitatively comparable results would be obtained under the weaker assumption that a spurious success is only detected with some probability greater than 0.5.

1. *For* $B' < \frac{-c\cdot(\alpha+\beta)}{\alpha}$*:* $p = 0$*,* $q_i = 0$*,* $\mu_A = \frac{\beta}{\beta+\alpha}$*,* $\mu_R = 1$*.*

   The probability that a published article is faulty is 0*. We call this equilibrium "Separation".*

2. *For* $B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}$ *and* $G < W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)\cdot\tau}$*:* $p = 1$*,* $q_i = 0$*,* $\mu_A = \frac{\beta}{\beta+\alpha}$*,* $\mu_R = \frac{\beta}{\beta+\alpha}$*.*

   The probability that a published article is faulty is $\frac{\alpha}{\beta+\alpha}$*, and the probability that a faulty article is revealed (if published) is 0. We call this equilibrium "Pooling I".*

3. *For* $B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}$*,* $G > W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)\cdot\tau}$ *and* $\tau < 1 - \sqrt[n]{\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}}$*:* $p = 1$*,* $q_i \in (0,1)$*,* $\mu_A = \frac{\beta}{\beta+\alpha}$*,* $\mu_R = \frac{\beta}{\beta+\alpha}$*.*

   Specifically, we have $q_i = \left(1 - \sqrt[n-1]{\frac{1}{W'}\cdot\left(G - \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)\cdot\tau}\right)}\right)\cdot\frac{1}{\tau}$*. The probability that a published article is faulty is* $\frac{\alpha}{\beta+\alpha}$*, and the probability that a faulty article is revealed (if published) is* $1-(1-q_i\cdot\tau)^n = 1 - \left(\frac{1}{W'}\cdot\left(G - \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)\cdot\tau}\right)\right)^{\frac{n}{n-1}}$*. We call this equilibrium "Pooling II".*

4. *For* $B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}$*,* $G \geq W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)\cdot\tau}$ *and* $\tau \geq 1 - \sqrt[n]{\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}}$*:* $p \in [0,1]$*,* $q_i \in [0,1]$*,* $\mu_A = \frac{\beta}{\beta+\alpha}$*,* $\mu_R = \frac{\beta}{\beta+\alpha\cdot p}$*.*

   Specifically, we have $p = \frac{\beta}{\alpha}\cdot\frac{k}{\tau\cdot\left(G - \left(\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}\right)^{\frac{n-1}{n}}\cdot W'\right) - k}$ *and* $q_i = \left(1 - \sqrt[n]{\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}}\right)\cdot\frac{1}{\tau}$*. The probability that a published article is faulty is* $\frac{\alpha}{\beta+\alpha}\cdot p$*, and the probability that a faulty article is detected (if published) is* $1-(1-q_i\cdot\tau)^n = 1 - \frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}$*. We call this equilibrium "Semi-Separation".*

Proof: See Appendix C.

Unlike in the deception game, we obtain a full separation equilibrium in which no false results are published, and therefore no reader ever wants to check a publication. This equilibrium occurs whenever the cost of investing care is sufficiently small as compared to

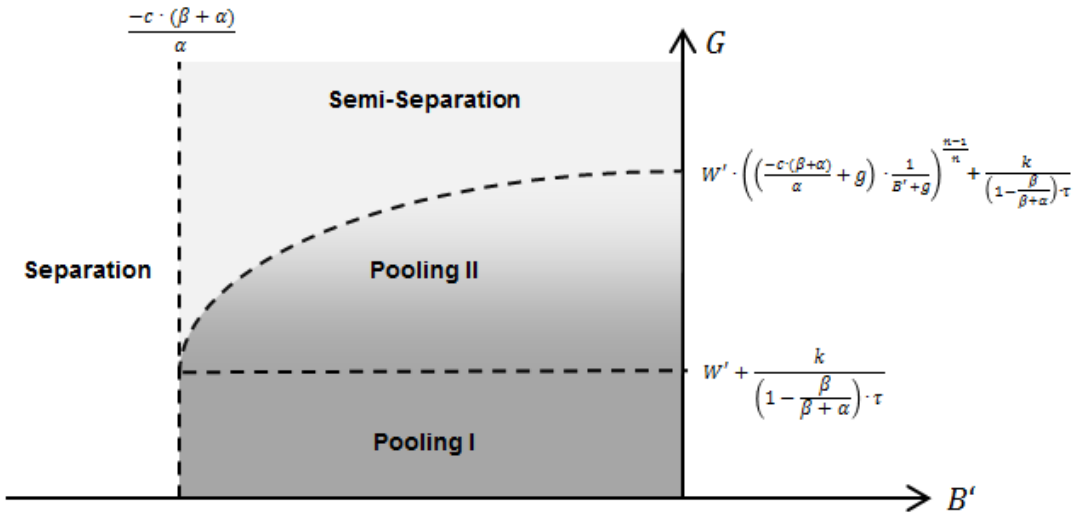

Figure 3.4: Resulting equilibria for different realizations of G and B'. Brighter shades refer to a higher overall probability of error detection.

the cost of mistakenly publishing an unsound finding (even if not detected) - that is, when $B' < \frac{-c\cdot(\alpha+\beta)}{\alpha}$. If this condition fails to hold, however, one of the remaining equilibria, all of which are qualitatively comparable to the deception game, will occur. Figure 3.4 illustrates all possible symmetric equilibria of the delusion game. The different (payoff) parameters have a similar effect on the occurrence of equilibria as in the deception game, and the above reasoning applies mutatis mutandis. Most importantly, Proposition 3.2 also applies to the delusion game, such that a larger readership entails a lower overall level of error detection in the "Pooling II" and "Semi-Separation" equilibria and a reduced frequency of diligence in the semi-separating equilibrium.

3.4 Extensions and Applications

3.4.1 The Priority Principle in Science

As mentioned earlier, the notion that a reader's (private) benefit G from successfully debunking fraudulent or erroneous research is independent of the number of other (successful) readers could be considered unrealistic. As soon as there are other readers that potentially check an article for fraud or errors, it seems reasonable that the expected benefit of successfully uncovering a deficient article is generally smaller. There are basically two lines of argument why this should be the case.

First, in academic research, priority matters (e.g. Dasgupta and David, 1994). If an

invalid publication is discovered by more than a single reader, there would always be the danger that a peer will outpace a successful reader and her scrutiny will be worthless in hindsight. From a reader's perspective, successfully debunking a flawed article alone is not enough. A reader must succeed in proving a publication's deficits, *and* she has to be the first one to do so. Second, one could alternatively suppose that all readers who have successfully revealed a flawed publication form a coalition and write a joint article. Then it is reasonable to assume that the reputational gain must be shared among all successful peers.

Either way, as soon as fellow readers check with positive probability, a reader's expected gain must be lower than in the setting in which G is independent of the number of successful readers. Therefore, we can derive the repercussions of the competition among readers from our reasoning on Proposition 3.1 and observe two effects. First, a reader's individual probability of checking an article will be smaller if an equilibrium of the type "Pooling II" emerges, since q_i is decreasing in G. Therefore, the overall probability of detection will likewise be lower in this kind of equilibrium. Second, in the semi-separating equilibrium, the author's propensity to cheat is higher, since p is increasing in G. Here, the individual checking probability is independent of G, and it is the author who responds to the level of G. Furthermore, it is easy to see that when G decreases, the set of parameter values that cause the second pooling equilibrium to emerge will expand at the expense of parameter that imply the existence of a semi-separating equilibrium.[25] Altogether we can conclude that the contest among readers further weakens the average quality of scientific publications. This result is in line with Kiri et al. (2018), who also find that the competition among scrutinizing scientists may undermine individual incentives to inspect scientific articles.

3.4.2 Heterogeneous Readers and Ideological Diversity

So far, our analysis has assumed that the entire audience consists of only homogeneous readers. This assumption is unrealistic, of course. In this section, we therefore add heterogeneity among the readers and focus on diversity in W', which measures a reader's negative payoff from a flawed article that passes unchallenged.[26] This payoff may vary

[25]Note that the set of parameter values that causes a "Pooling I"-equilibrium to emerge will not increase, since the expected size of G is only affected if $q_i > 0$.

[26]We concentrate on the deception game, but similar results are obtainable for the delusion game.

for different readers due to their differences in scientific or ideological standpoints. A mistakenly accepted article could be less troublesome for a reader whose own theory or worldview is supported by the wrong result than for her colleague whose own theory is (erroneously) rebutted by the publication. Therefore, W' can act as a measure of a reader's (ideological) distance from the author's position. As discussed in section 3.3.1, if W' decreases for *all* readers, the total level of fraud decreases and the level of detected fraud weakly increases. This means if the entire readership were more opposed to the author's position, there would be a greater probability of keeping science clean.

What is less straightforward to see is whether *diversity* of readers is conducive to increasing the quality of science. It has been argued that heterogeneity within deliberating groups is beneficial and helps to come closer to the truth (e.g. Surowiecki 2004, Page 2006). Here, we concentrate on the simplified case of $n = 2$ readers. In particular, we are interested in the question whether a readership with payoffs $W_1' = W' + \Delta$ and $W_2' = W' - \Delta$ is superior in terms of fraud (detection) volumes to a homogeneous readership with $W_1' = W_2' = W'$. We restrict our analysis to the case of $G \geq W' + \frac{k}{(1-\beta)}$ and $\tau = 1$. These conditions ensure the existence of the semi-separating equilibrium for an audience consisting of symmetric readers. In the benchmark case with completely symmetric readers, the respective probabilities of fraud and inspection are $p = \frac{\beta}{1-\beta} \cdot \frac{k}{G-W' \cdot \sqrt{\frac{g}{B'+g}}-k}$ and $q_i = 1 - \sqrt{\frac{g}{B'+g}}$. This symmetric equilibrium no longer exists when we have asymmetric readers:

Lemma 3.1. *There is no symmetric equilibrium when* $W_1' = W' + \Delta$ *and* $W_1' = W' - \Delta$.

Proof: The respective indifference conditions for readers 1 and 2 are

$$\mu \cdot W + (1-\mu) \cdot G - k = \mu \cdot W + (1-\mu) \cdot (1-q_2) \cdot (W' + \Delta) \tag{3.4}$$

and

$$\mu \cdot W + (1-\mu) \cdot G - k = \mu \cdot W + (1-\mu) \cdot (1-q_1) \cdot (W' - \Delta). \tag{3.5}$$

Evidently, $q_1 = q_2$ is only possible if $\Delta = 0$. □

We compare fraud and detection frequencies in the symmetric benchmark case to those of two distinct equilibria which emerge when readers are not symmetric. It turns out that a heterogeneous audience *potentially* reduces the level of fraud, i.e. there is always

an equilibrium in which the author is less inclined to cheat. That said, reader heterogeneity is *not* sufficient for lower fraud levels. For some parameter constellations, there exists a further equilibrium in which fraud levels are higher in the case of an asymmetric readership.

Proposition 3.4. *In the deception game with two heterogeneous readers ($W_1' = W' + \Delta$ and $W_2' = W' - \Delta$) and $\tau = 1$, there exist equilibria such that*

1. *For $G \geq W' + \Delta + \frac{k}{1-\beta}$ and $\Delta < W' \cdot \left(\frac{-B'}{B'+2 \cdot g}\right)$: $p = \frac{\beta}{1-\beta} \cdot \frac{k}{G-(W'+\Delta)-k}$, $q_1 = 1 - \frac{g}{B'+g}$ and $q_2 = 0$.*

   In this equilibrium, compared to the symmetric benchmark equilibrium ($W_1' = W_2' = W'$), the author's probability of cheating is strictly higher. The overall probability of detecting existing fraud is identical in this equilibrium and in the benchmark equilibrium.

2. *For $G \geq W' + \frac{k}{1-\beta}$: $p = \frac{\beta}{1-\beta} \cdot \frac{k}{G-(W'-\Delta)-k}$, $q_1 = 0$ and $q_2 = 1 - \frac{g}{B'+g}$.*

   In this equilibrium, compared to the symmetric benchmark equilibrium ($W_1' = W_2' = W'$), the author's probability of cheating is strictly lower. The overall probability of detecting existing fraud is identical in this equilibrium and in the benchmark equilibrium.

Proof: See Appendix C.

Both equilibria are semi-separating equilibria.[27] In each of them, one reader (whom we will refer to as "active") checks a publication with positive probability while the other reader remains completely idle. When it is the second reader - the one who suffers more from an unpunished deceptive article - who possibly checks whether a published article is fraudulent, the author's volume of fraud is distinctly lower than in the symmetric equilibrium with two homogeneous readers. The respective values of p and q_2 are identical to those that we would obtain, would there be only a single reader with $W_1' = W' - \Delta$.

[27] For a restricted parameter range, there exists one further semi-separating equilibrium in which both readers check with positive and dissimilar probabilities. We do not analyze this equilibrium in any greater detail here and concentrate instead on the most asymmetric ones.

Within our restricted parameter space, this equilibrium always exists. Hence, a more heterogeneous audience might result in lower fraud levels. All the same, if Δ is "small", there exists another equilibrium in which it is the first reader who potentially scrutinizes a publication, whereas the second reader remains completely inactive. This might be somewhat surprising, since the first reader, compared to his inactive peer, profits *less* from a successfully debunked article. Consequently, comparing the two asymmetric equilibria, the level of fraud is higher when only the first reader is active, as the author must cheat more often to make her indifferent between checking and not checking.
Comparing this equilibrium to the symmetric equilibrium with homogeneous readers, it turns out that the author's probability of cheating will be higher in the case of a single active reader. Therefore, we can conclude that heterogeneity (as defined here) is not sufficient to unambiguously reduce the volume of fraud in science. We can furthermore conclude that the volume of *observed* fraud will be highest if only reader 1 is active, lowest if only reader 2 is active, and between these extremes if we have a homogeneous audience that forms a symmetric equilibrium. This is directly implied by the fact that upon cheating, the probability of being discovered is identical in all three equilibria.

3.4.3 Strategic Audience Choice

A rational (and malevolent) author will be aware that the free rider problem is more pronounced in the presence of larger audience sizes. Therefore, all other parameters held constant, he will always weakly prefer a huge audience over a small one. In fact, as we will show in this section, the author can even have incentives to *strategically induce* the free rider problem, i.e. to take measures that are suited to increasing the number of interested readers. The most obvious way to attract a higher number of readers to a scientific article is by offering a more interesting or spectacular result. If we allow for the possibility of more than two output levels, then this can have quite severe implications for the level of undetected fraud.
In the following, we modify the deception game at stage 1 slightly and allow for a third outcome L that represents a landmark result which is generally suited to arousing the interest of a larger audience. Such a breakthrough occurs with probability β^L, whereas a success occurs with probability β^S. At stage 2, if A experiences a failure, he can now choose between three options. Besides (*no pub* F) and (*pub* $\widetilde{S}$), he can also choose

to oversell the failure as a landmark result ($pub\ \widetilde{L}$).[28] We will refer to the different transgression levels as "mild cheating" and "heavy cheating" and $p^{\widetilde{S}}$ and $p^{\widetilde{L}}$ denote the respective playing probabilities.[29]

Articles that contain an (alleged) landmark result will attract a larger audience than those that "only" present a success, i.e. $n^L > n^S$ and q_i^Y denotes the individual checking probability for a member of audience Y. In the absence of further checking readers, a member of audience Y wants to check an article if

$$G^Y > W'^Y + \frac{k^Y}{\left(1 - \frac{\beta^Y}{\beta^Y + (1-\beta^L-\beta^S)}\right) \cdot \tau}. \tag{3.6}$$

We assume here that this assumption is always satisfied. In order to further simplify the analysis and concentrate on the most interesting result, we also rule out semi-separating equilibria in this section by assuming that

$$(1-\tau)^{n^Y} > \frac{g^Y}{B'^Y + g^Y} \tag{3.7}$$

for $Y = L, S$. It is furthermore reasonable to assume $G^L \geq G^S$, $W'^L \leq W'^S = W'$ and $k^L = k^S = k$, though these assumptions are not crucial for our qualitative results. Thus, the rebuttal of a more spectacular research result will yield a weakly higher benefit for any reader. On the other hand, any reader will obtain a weakly higher loss when the result is mistakenly accepted by the scientific community.

For the author's payoff parameters, it is straightforward to assume that $B'^L \geq B'^S$ and $g^L \geq g^S$. Both the reward (for unrevealed fraud) and the punishment (for revealed fraud) are higher if the author claims to have produced a landmark result instead of a success.

The game's symmetric equilibria are characterized in the following proposition:

Proposition 3.5. *For the deception game with two transgression levels, "mild cheating" and "heavy cheating", two audiences ($n^L > n^S$), and assuming that $G^Y > W'^Y + \frac{k}{(1-\beta^L-\beta^S)\cdot\tau}$, $(1-\tau)^{n^Y} > \frac{g^Y}{B'^Y+g^Y}$, every parameter constellation yields exactly one symmetric equilibrium, such that*

[28] We exclude the possibility of overselling a success as a landmark result and assume that the author always publishes a success as such.

[29] Gall and Maniadis (2019) discuss different types of questionable research practices and distinguish between rather mild and more severe forms of scientific misconduct.

1. *For* $\left(\frac{1}{W'^S}\cdot\left(G^S-\frac{k}{\left(1-\frac{\beta^S}{1-\beta^L}\right)\cdot\tau}\right)\right)^{\frac{n^S}{n^S-1}}\cdot(B'^S+g^S)-g^S<\left(\frac{1}{W'^L}\cdot\left(G^L-\frac{k}{\left(1-\frac{\beta^L}{1-\beta^S}\right)\cdot\tau}\right)\right)^{\frac{n^L}{n^L-1}}\cdot(B'^L+g^L)-g^L$*:* $q_i^L\in(0,1)$, $p^{\widetilde{L}}=1$, $p^{\widetilde{S}}=0$, $\mu^L=\frac{\beta^L}{1-\beta^S}$, $\mu_S=\beta^S$.

   Specifically, we have $q_i^L=\left(1-\sqrt[n^L-1]{\frac{1}{W'}\cdot\left(G-\frac{k}{\left(1-\frac{\beta^S}{1-\beta^L}\right)\cdot\tau}\right)}\right)\cdot\frac{1}{\tau}$.

2. *For* $\left(\frac{1}{W'^S}\cdot\left(G^S-\frac{k}{\left(1-\frac{\beta^S}{1-\beta^L}\right)\cdot\tau}\right)\right)^{\frac{n^S}{n^S-1}}\cdot(B'^S+g^S)-g^S\geq\left(\frac{1}{W'^L}\cdot\left(G^L-\frac{k}{\left(1-\frac{\beta^L}{1-\beta^S}\right)\cdot\tau}\right)\right)^{\frac{n^L}{n^L-1}}\cdot(B'^L+g^L)-g^L$*:* $q_i^S\in(0,1)$, $p^{\widetilde{L}}=0$, $p^{\widetilde{S}}=1$, $\mu^L=\beta^L$, $\mu_S=\frac{\beta^S}{1-\beta^L}$.

   Specifically, we have $q_i^S=\left(1-\sqrt[n^S-1]{\frac{1}{W'}\cdot\left(G-\frac{k}{\left(1-\frac{\beta^L}{1-\beta^S}\right)\cdot\tau}\right)}\right)\cdot\frac{1}{\tau}$.

Should the author cheat heavily, his risk of getting caught is strictly lower compared to mild cheating if $\left(\frac{1}{W'^L}\cdot\left(G^L-\frac{k}{\left(1-\frac{\beta^L}{1-\beta^S}\right)\cdot\tau}\right)\right)^{\frac{n^L}{n^L-1}}<\left(\frac{1}{W'^S}\cdot\left(G^S-\frac{k}{\left(1-\frac{\beta^S}{1-\beta^L}\right)\cdot\tau}\right)\right)^{\frac{n^S}{n^S-1}}$.

Proof: See Appendix C.

From a perspective of optimal incentive design, these results are somewhat discouraging. By committing a greater offense (heavy cheating), the perpetrator can leave himself better off and actually *reduce* the probability of getting caught. Still, an author might prefer a milder transgression level if a more severe punishment for heavy cheating offsets the reduced likelihood of getting caught.

The above results are also interesting in light of the findings of Furman et al. (2012) and their discussion in L&Z. In the case of biomedicine, the former authors find that it is mostly highly influential research that is retracted after publication and that retractions are relatively scarce in low-profile research. L&Z speculate that this finding can be explained by a reader's low reward for refuting "incremental" research, which is therefore not scrutinized at all. In line with the above findings, our alternative explanation for this phenomenon would be that a researcher only commits fraud if doing so earns him a high-profile publication and therefore attracts many readers. We therefore might not see much low-profile fraudulent research, not because it remains undetected, but because it rarely exists.

3.4.4 Editors and Peer Review

The process of peer review as a central element of academic publishing has been neglected in our analysis so far. In this section, we therefore add a (male) journal editor (E) to the set of players. He has the possibility to check an article *before* its publication and can condition the publication decision on the outcome of the review process.[30]

Since reviewers are supposed to check for errors and shortcomings rather than for outright fraud, we focus on the delusion game and analyze how equilibria are affected by the presence of the editor. In particular, the baseline model is adjusted as follows:

We assume that there is only *one* reader who finds errors - if present - with certainty, i.e. $\tau_R = 1$. Stages 1 and 2 are identical to the original game. At stage 3, after A has made his decision whether or not to invest care, he decides to submit or not submit the resulting article. The payoffs are such that the author will always submit an uncertain and a certain success, but will never submit a certain failure. At stage 4, the editor observes the author's submission decision and updates μ_E, which denotes the probability that a flawless article has been produced. He then decides whether to simply rubber-stamp the submission without closer inspection (*no review* played with probability r) or to conduct a proper review (*review* played with probability $1 - r$). Should E decide to conduct a proper review, he must invest $k_E > 0$ and finds existing errors with $\tau_E = 1$. If the paper is found to be clean, it gets published and the game enters stage 5. Otherwise, when errors are found, the submission is rejected and the author obtains a utility loss f because of the failed submission. We assume that $g > f \geq 0$, such that the author prefers a flawed publication to be detected by the editor instead of an alert reader. Should E decide not to conduct a proper review, the paper will be published in any case and the game goes straight to stage 5. For the reader, it is unknown whether the author has invested care and also whether the editor has conducted a thorough review. Like in the original game, R decides whether she wants to scrutinize the publication. Payoffs for author and reader are analogous to the original game. We assume that the editor obtains a negative payoff B'_E if a flawed article clears the hurdle of peer review but remains undetected by the reader. Should the reader instead reveal the article to be erroneous, the editor will incur

[30] Usually, the editor will not check a submission himself but delegates this task to one or several reviewers. We abstract from this principal-agent problem and assume that the editor performs the check himself.

an even higher loss g_E. If the article is clean, the editor gains a positive benefit $B_E > 0$. The most important differences to the (extended) model of L&Z is that, unlike them, we model the editor's behavior as endogenous and also draw conclusions about the author's (changed) behavior when an editor is involved. Moreover, in our model, for the author we regard the rejection by the editor and the rebuttal by the scientific community as two distinct events. We restrict our attention to the most interesting case where $B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}$ and $G \geq W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)}$. These assumptions imply the emergence of a semi-separating equilibrium for the original game in which no editor is present. In this case, A does not invest care with probability $p = \frac{\beta}{\alpha} \cdot \frac{k}{G-W'-k}$ and R checks a publication with probability $q = 1 - \frac{(-c+g)\cdot\alpha - c\cdot\beta}{(B'+g)\cdot\alpha}$.

In the following proposition, we show that the peer review process can have adverse effects on the average quality of published articlessince the presence of an editor might crowd out the author's own incentive to thoroughly scrutinize an article before it gets published. In particular, there exists an equilibrium that leaves the overall volume of erroneous research unchanged, but reduces the likelihood that a faulty article will be revealed by the reader.

Proposition 3.6. *For the delusion game, with one reader and one editor and* $\tau_E = \tau_R = 1$, *there exists an equilibrium such that for* $f < \frac{\frac{c\cdot(\alpha+\beta)}{\alpha} + \frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}\cdot\left(\left(1-\frac{(-k_E+g_E)\cdot\alpha-k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha}\right)\cdot(-g-B')+B'\right)}{1-\frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}}$, $B'_E \geq \frac{-k_E\cdot(\alpha+\beta)}{\alpha}$ *and* $G \geq W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)}$*:* $p = 0$, $q = 1 - \frac{(-k_E+g_E)\cdot\alpha - k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha}$, $r = \frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}$, $\mu_A = \frac{\beta}{\beta+\alpha}$, $\mu_E = \frac{\beta}{\beta+\alpha}$, $\mu_R = \frac{\beta}{\beta+\alpha\cdot r}$.

The probability that a published article is faulty is $\frac{\beta}{\beta+\alpha}\cdot\frac{k}{G-W'-k}$ *and equal to the respective probability of the original game without an editor. The probability that a faulty article gets detected (if published) is* $1 - \frac{(-k_E+g_E)\cdot\alpha - k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha}$ *and lower than in the original game whenever* $\frac{B'_E+g_E}{B'+g} > \frac{(-k_E+g_E)\cdot\alpha - k_E\cdot\beta}{(-c+g)\cdot\alpha - c\cdot\beta}$.

Proof: See Appendix C.

In the above equilibrium, it is the editor that (sometimes) checks an article for flaws while the author remains idle. This situation is likely to occur when f (the author's utility loss from being rejected by the editor) is small. Since the editor might lose his reputation if he erroneously accepts flawed publications, he will (at least sometimes) check publications for errors. Now it is he who makes the reader indifferent between checking and not checking

a publication. Therefore, the probability of a publication being erroneous is $\frac{\beta}{\beta+\alpha} \cdot \frac{k}{G-W'-k}$ in both versions of the game. Whether q is higher or lower than in the original game depends on the ratio of the author's and the editor's payoff parameters.

The existence of this equilibrium shows that the presence of an editor (or more generally the process of peer review) does not necessarily lead to a lower error rate in scientific articles and that the volume of errors revealed by readers can even decrease.[31] The game also illustrates that the author's and editor's interests *after* the release of an article are aligned, since they both hope not to be debunked by an alert reader.

3.5 Discussion

In this section, we want to briefly discuss the above findings. First, it is worthwhile highlighting the differences between our model and those of L&Z and Kiri et al. (2018). While L&Z mainly focus on different types of research (incremental vs. radical) and their respective odds of being fraudulently produced (and of being revealed as such), we disregard this distinction between research types and focus instead on the scientific community and its role in debunking deficient publications. Kiri et al. (2018) concentrate on the author's motivation for investing costly effort that positively affects the chances of producing high-quality research when the resulting article will possibly undergo a check by a single colleague. In an extension of their model, they introduce a second reader who can also check a publication's validity. Like us, they find that the overall probability of debunking low-quality research can decrease by introducing an additional reader. However, the mechanism at work is completely different from ours since their results are driven solely by the readers' quest for priority. Our main contribution is therefore to show that a volunteer's dilemma exists among members of the scientific community and to analyze how this dilemma depends on the size of the relevant community. We also study how this dilemma in turn influences the author's willingness to cheat or to apply an insufficient level of diligence.

Our central finding is certainly Proposition 3.2, in which we show that an increasing number of readers is possibly detrimental to the average quality of scientific publications. This is clearly a counter-intuitive finding. What can positively affect the volume of detected flawed research, though, is a high level of G (or equivalently a low level of

[31] The presented result would also hold in the presence of $n > 1$ readers.

W'). This shows that for the scientific community to be self-correcting, it is crucial that readers have some (ideological) distance from the author's presented findings. Our results therefore suggest that an audience consisting of "devil's advocates", i.e. readers who are committed to a different theory or paradigm, is certainly helpful for reducing the volume of flawed publications. Certainly, our central finding can be challenged for several reasons. First, as already stated, the payoff parameters of all players are most likely not independent of the number of readers. Second, our model does not explicitly address the role of follow-up research. More interesting findings (those with many readers) are more likely to spur future research activities, which could be helpful for refuting unsound articles. Third, the readers' individual probability of finding errors might be not independent of each other, but positively or negatively correlated.

There are some avenues for future research. First, it is not entirely clear to what degree our results hold more generally when we allow for richer action spaces, e.g. a continuum of checking levels or cheating levels. Second, as a possible extension, one could explicitly incorporate the rivalry among authors competing for scarce journal space (similar to Gall and Maniadis (2019)). The implications might be different from those of L&Z, who model a harsher "publish or perish" paradigm simply by having a higher individual publication benefit. Third, our finding that the frequency of deception increases with the number of potential law enforcers might be relevant in contexts other than academic publishing.[32]

3.6 Conclusion

We have presented a model of the scientific approval process where scrutinizing scientific publications is individually costly and causes a positive externality for the whole scientific community. In the model's basic version, an author can decide to publish a fraudulent article if a research project turns out to be a failure. Contrary to the intuitive view that a higher number of readers should be more effective at deterring authors from behaving fraudulently and also increase the number of detected cases of fraud, we find that the contrary might be true, depending on parameter size. The effect is due to the readers' individual free riding behavior, which in turn affects the author's willingness to cheat. Therefore, our model challenges the notion of self-correcting science. In an ad-

[32]Think of a politician who wants to cheat a large electorate or an agent who wants to deceive a collective of principals.

justed version of the model, we have analyzed the case of erroneous research, where a flawed publication is the result of a lack of diligence, rather than deliberate fraud. Likewise, increasing readership size might be detrimental rather than conducive to reducing and uncovering deficient publications. Incorporating the readers' competition for priority might boost the level of (undetected) defective research further. If we explicitly consider the possibility of two transgression levels (mild and severe misconduct), it turns out that an author who opts for the severe transgression level can actually reduce his risk of getting caught because the free rider problem is more pronounced in the presence of severe misconduct. Moreover, neither greater levels of ideological diversity among the readership nor the presence of a peer review process unambigiously reduce the volume of deficient research.

C Appendix

Proofs

Proof of Proposition 3.1

We start by showing that for all possible equilibria, the author will always publish an article if he gains a success. Publishing a successful project is superior to not publishing if

$$(1-(1-q_i\cdot\tau)^n)\cdot B+(1-q_i\cdot\tau)^n\cdot B\geq 0\Leftrightarrow B\geq 0. \tag{C.1}$$

This is true by assumption.
The condition that makes any reader prefer to not check a publication is

$$\begin{aligned}\mu\cdot W+(1-\mu)\cdot\left(\tau\cdot G+(1-\tau)\cdot(1-q_i\cdot\tau)^n\cdot W'\right)-k<\\ \mu\cdot W+(1-\mu)\cdot\left(1-q_i\cdot\tau\right)^{n-1}\cdot W'\end{aligned} \tag{C.2}$$

and $\mu=P(S|article)=\frac{P(article|S)\cdot P(S)}{P(article|S)\cdot P(S)+P(article|F)\cdot P(F)}=\frac{\beta}{\beta+p\cdot(1-\beta)}$.
The author's condition for cheating ($pub\ \widetilde{S}$) to be rational and to set $p=1$ is

$$(1-(1-q_i\cdot\tau)^n)\cdot(-g)+(1-q_i\cdot\tau)^n\cdot B'>0\Leftrightarrow(1-q_i\cdot\tau)^n>\frac{g}{B'+g}. \tag{C.3}$$

For "Pooling I" to exist, condition (C.3) must be satisfied with $q_i=0$, such that the condition degenerates to $B'>0$, a condition that is always true. Then, $\mu=\beta$ and condition (C.2) yields

$$G<W'+\frac{k}{(1-\beta)\cdot\tau}. \tag{C.4}$$

One can readily see that parameters that meet this condition can be easily found.
For "Pooling II" to exist, it must be that (C.3) is strictly satisfied. This implies that $\mu=\beta$. Since the author will always decide to publish a paper, the readers' updated posterior will be identical to the prior. The publication decision is not informative with respect to the research project's outcome (success or failure). Furthermore, condition (C.2) must hold with equality. We then get

$$q_i=\left(1-\sqrt[n-1]{\frac{1}{W'}\cdot\left(G-\frac{k}{(1-\beta)\cdot\tau}\right)}\right)\cdot\frac{1}{\tau}. \tag{C.5}$$

To obtain $q_i\in(0,1)$, the following conditions must be satisfied: We have

$$q_i>0\Leftrightarrow G>W'+\frac{k}{(1-\beta)\cdot\tau} \tag{C.6}$$

and

$$q_i < 1 \Leftrightarrow G < W' \cdot (1-\tau)^{n-1} + \frac{k}{(1-\beta)\cdot\tau}. \tag{C.7}$$

Condition (C.7) is identical to condition (3.3) and true by assumption.
Substituting (C.5) into (C.3) yields

$$\sqrt[n-1]{\frac{1}{W'}\cdot\left(G - \frac{k}{(1-\beta)\cdot\tau}\right)} > \sqrt[n]{\frac{g}{B'+g}} \Leftrightarrow G < W' \cdot \left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}} + \frac{k}{(1-\beta)\cdot\tau}. \tag{C.8}$$

The above inequality is implied by (3.3) if

$$\left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}} < (1-\tau)^{n-1} \Leftrightarrow \frac{g}{B'+g} < (1-\tau)^n. \tag{C.9}$$

It is easy to see that the set of parameters, defined by (C.6), (C.7) and (C.9), is non-empty for any $\tau \in (0,1)$.
We proceed with "Semi-Separation". Inequality (C.3) must hold with equality, and we obtain

$$\Leftrightarrow q_i = \left(1 - \sqrt[n]{\frac{g}{B'+g}}\right)\cdot\frac{1}{\tau}. \tag{C.10}$$

Inequality (C.2) must also hold with equality and yields

$$\tau\cdot\left(G - (1-q_i\cdot\tau)^{n-1}\cdot W'\right) - \frac{k}{1-\mu} = 0. \tag{C.11}$$

We can solve for p and obtain

$$\begin{aligned} p &= \frac{\beta}{1-\beta}\cdot\frac{k}{\tau\cdot\left(G - W'\cdot(1-q_i\cdot\tau)^{n-1}\right) - k} \\ &\Leftrightarrow \frac{\beta}{1-\beta}\cdot\frac{k}{\tau\cdot\left(G - W'\cdot\left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}}\right) - k}. \end{aligned} \tag{C.12}$$

Next, we derive the conditions for which $q_i, p \in [0,1]$. For q_i we obtain

$$q_i \geq 0 \Leftrightarrow 1 \geq \frac{g}{B'+g} \Leftrightarrow B' \geq 0 \tag{C.13}$$

and

$$q_i \leq 1 \Leftrightarrow (1-\tau)^n \leq \frac{g}{B'+g} \tag{C.14}$$

and the first condition is always true. We have furthermore

$$p \geq 0 \Leftrightarrow \frac{\beta}{1-\beta} \cdot \frac{k}{\tau \cdot \left(G - W' \cdot (1 - q_i \cdot \tau)^{n-1}\right) - k} \geq 0$$
$$\Leftrightarrow G \geq W' \cdot (1 - q_i \cdot \tau)^{n-1} + \frac{k}{\tau} \tag{C.15}$$

and

$$p \leq 1 \Leftrightarrow \frac{\beta}{1-\beta} \cdot \frac{k}{\tau \cdot \left(G - W' \cdot (1 - q_i \cdot \tau)^{n-1}\right) - k} \leq 1$$
$$\Leftrightarrow G \geq W' \cdot (1 - q_i \cdot \tau)^{n-1} + \frac{k}{(1-\beta) \cdot \tau}. \tag{C.16}$$

Inequality (C.15) is less restrictive than inequality (C.16) and is therefore not binding. Plugging (C.10) into (C.16) yields

$$G \geq W' \cdot \left(\frac{g}{B' + g}\right)^{\frac{n-1}{n}} + \frac{k}{(1-\beta) \cdot \tau}. \tag{C.17}$$

Referring to condition (3.3), it is easy to see that condition (C.17) holds whenever conditions (C.6) and (C.14) are satisfied. □

Proof of Proposition 3.2

In the first pooling equilibrium, n readers prefer to not check a publication. It is straightforward to see that if n readers prefer to remain idle, this is also true for $n+1$ readers since the validity of inequality (C.4) remains unaffected by the additional reader. The volume of fraudulent research remains $(1-\beta)$ and no bogus article will ever get revealed. If "Pooling II" exists for n readers, there will always exist such an equilibrium for $n+1$ readers. This is the case because the conditions for its existence remain either unaffected (condition (C.6)), or become weaker (condition (C.9)) if the number of readers is increased. Making use of equation (C.5), the difference in the overall levels of undetected fraud equals

$$\left(1 - q_i^{n+1} \cdot \tau\right)^{n+1} - \left(1 - q_i^{n} \cdot \tau\right)^{n}, \tag{C.18}$$

where q_i^{n+1} and q_i^n respectively denote individual checking probabilities for $n+1$ and n readers. Plugging in the respective individual probabilities, we obtain

$$\left(\frac{1}{W'} \cdot \left(G - \frac{k}{(1-\beta) \cdot \tau}\right)\right)^{\frac{n+1}{n}} - \left(\frac{1}{W'} \cdot \left(G - \frac{k}{(1-\beta) \cdot \tau}\right)\right)^{\frac{n}{n-1}}. \tag{C.19}$$

The base of both sides must range in the closed unit interval (otherwise q_i could not be

$\in (0,1)$), and therefore the difference is positive when $\frac{n+1}{n} < \frac{n}{n-1} \Leftrightarrow n^2 > n^2 - 1$, which is obviously true.

When the parameter constellation for n readers implies the existence of a semi-separating equilibrium, we have to distinguish two cases: In case 1, if $(1-\tau)^{n+1} \geq \frac{g}{B'+g}$, a semi-separating equilibrium will also occur for $n+1$ readers. Then, referring to equation (C.12), the difference in fraud levels with $n+1$ and n readers respectively equals

$$\begin{aligned} &\frac{\beta}{1-\beta} \cdot \frac{k}{\tau \cdot \left(G - W' \cdot \left(\frac{g}{B'+g}\right)^{\frac{n}{n+1}}\right) - k} - \frac{\beta}{1-\beta} \cdot \frac{k}{\tau \cdot \left(G - W' \cdot \left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}}\right) - k} \\ \Leftrightarrow\ &\frac{\beta}{1-\beta} \cdot \left(\frac{k}{\tau \cdot \left(G - W' \cdot \left(\frac{g}{B'+g}\right)^{\frac{n}{n+1}}\right) - k} - \frac{k}{\tau \cdot \left(G - W' \cdot \left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}}\right) - k} \right). \end{aligned} \tag{C.20}$$

As one can easily see, the term is positive whenever

$$\left(\frac{g}{B'+g}\right)^{\frac{n}{n+1}} < \left(\frac{g}{B'+g}\right)^{\frac{n-1}{n}} \Leftrightarrow n^2 - 1 < n^2. \tag{C.21}$$

Making use of equation (C.10), we see that after publication has been released, the fraction of articles that get scrutinized is not affected by n since

$$\begin{aligned} 1 - \left(1 - \left(1 - \sqrt[n]{\frac{g}{B'+g}}\right)\right)^n = 1 - \left(1 - \left(1 - \sqrt[n+1]{\frac{g}{B'+g}}\right)\right)^{n+1} \\ \Leftrightarrow 1 - \frac{g}{B'+g} = 1 - \frac{g}{B'+g}. \end{aligned} \tag{C.22}$$

Since the overall volume of fraud increases and the share of debunked fraudulent articles remains constant, the absolute volume of undetected fraudulent articles is higher for $n+1$ readers than for n readers.

In case 2, if $(1-\tau)^{n+1} \geq \frac{g}{B'+g}$, a pooling equilibrium will emerge for $n+1$ readers. Hence, A now strictly prefers cheating and is not kept indifferent between cheating and not publishing any longer. Since A's payoff only depends on the overall likelihood of getting caught, we know that this likelihood is smaller in the pooling equilibrium than in the semi-separating equilibrium.

It is obvious that the parameter set for which "Pooling II" is an equilibrium weakly expands (and the set for which the semi-separating equilibrium exists weakly decreases) since $(1-\tau)^{n+1} \leq (1-\tau)^n$. □

Proof of Proposition 3.3
We start with "Separation". The author will set $p = 0$ and always invest care if

$$\begin{aligned} \mu_A \cdot B + (1 - \mu_A) \cdot 0 - c > \\ \mu_A \cdot B + \left(1 - \mu_A^1\right) \cdot \left((1 - (1 - q_i \cdot \tau)^n) \cdot (-g) + (1 - q_i \cdot \tau)^n \cdot B'\right) \\ \Leftrightarrow \frac{-c \cdot (\alpha + \beta)}{\alpha} > (1 - q_i \cdot \tau)^n \cdot \left(B' + g\right) - g \end{aligned} \tag{C.23}$$

and $\mu_A = P(true\ success|observed\ success) = \frac{P(observed\ success|true\ success) \cdot P(true\ success)}{P(observed\ success)} = \frac{\beta}{\beta+\alpha}$.
Furthermore, any reader will abstain from checking a publication and set $q_i = 0$ if

$$\begin{aligned} \mu_R \cdot W + (1 - \mu_R) \cdot \left(\tau \cdot G + (1 - \tau) \cdot (1 - q_i \cdot \tau)^{n-1} \cdot W'\right) - k < \\ \mu_R \cdot W + (1 - \mu_R) \cdot (1 - q_i \cdot \tau)^{n-1} \cdot W' \end{aligned} \tag{C.24}$$

and $\mu_R = P(true\ success|article) = \frac{P(article|true\ success) \cdot P(true\ success)}{P(article)} = \frac{\beta}{\beta+\alpha \cdot p}$.
For the equilibrium to exist, it must be that inequalities (C.23) and (C.24) both strictly hold for $p = 0$ and $q_i = 0$. Then, $\mu_R = 1$ and condition (C.24) reduces to $W - k < W$, which is always true. For $q_i = 0$, condition (C.23) can be simplified to

$$B' < \frac{-c \cdot (\alpha + \beta)}{\alpha}. \tag{C.25}$$

Therefore, condition (C.25) alone is sufficient for the postulated equilibrium to exist.
Next, we prove the existence of the second pooling equilibrium ("Pooling II"). First, condition (C.23) must hold with reversed operator, such that "no care" yields the author a weakly higher utility than "care" and he sets $p = 1$. This implies that $\mu_R = \frac{\beta}{\beta+\alpha}$. From condition (C.23) we can then conclude that

$$(1 - q_i \cdot \tau)^n \geq \frac{(-c + g) \cdot \alpha - c \cdot \beta}{\alpha \cdot (B' + g)}. \tag{C.26}$$

Readers mix between checking and not checking, and condition (C.24) holds with equality. We then obtain

$$q_i = \left(1 - \sqrt[n-1]{\frac{1}{W'} \cdot \left(G - \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}\right)}\right) \cdot \frac{1}{\tau}. \tag{C.27}$$

The conditions that ensure that $q \in (0, 1)$ yield

$$q_i > 0 \Leftrightarrow G > W' + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}, \tag{C.28}$$

as well as

$$q_i < 1 \Leftrightarrow G < W' \cdot (1-\tau)^{n-1} + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}. \tag{C.29}$$

This last condition is true by assumption.

Substituting (C.27) into (C.26) yields

$$G < W' \cdot \left(\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}\right)^{\frac{n-1}{n}} + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}. \tag{C.30}$$

The above inequality is implied by (C.29) if

$$\begin{aligned}\left(\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}\right)^{\frac{n-1}{n}} < (1-\tau)^{n-1} &\Leftrightarrow \frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)} < (1-\tau)^{n} \\ &\Leftrightarrow \tau < 1 - \sqrt[n]{\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}}\end{aligned} \tag{C.31}$$

One can easily verify that inequalities (C.28) and (C.31) can simultaneously hold true for any $\tau \in (0,1)$. Moreover, $B' \geq \frac{-c(\alpha+\beta)}{\alpha}$ is implied if (C.31) holds true.

We proceed with the first pooling equilibrium. The readers will not check any publication ($q_i = 0$) if (C.25) holds true, again with $\mu_R = \frac{\beta}{\beta+\alpha}$. Then, the inequality can be transformed to

$$G < W' + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}. \tag{C.32}$$

Referring to condition (C.23), the author prefers not to invest care and to set $p = 1$ if

$$B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}. \tag{C.33}$$

It is straightforward to see that conditions (C.32) and (C.33) can hold simultaneously for a non-empty set of parameter values.

Finally, we prove the existence of the semi-separating equilibrium. The readers set q_i such that A is indifferent between "no care" and "care", and condition (C.23) must hold with equality and yields

$$q_i = \left(1 - \sqrt[n]{\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}}\right) \cdot \frac{1}{\tau}. \tag{C.34}$$

Likewise, the author makes all readers indifferent between "check" and "no check", and (C.24) holds with equality:

$$\tau \cdot \Big(G - (1 - q_i \cdot \tau)^{n-1} \cdot W'\Big) - \frac{k}{1 - \mu_R} = 0 \tag{C.35}$$

and $\mu_R = \frac{\beta}{\beta + \alpha \cdot p}$. We solve for p and obtain

$$\begin{aligned} p &= \frac{\beta}{\alpha} \cdot \frac{k}{\tau \cdot \Big(G - (1 - q_i \cdot \tau)^{n-1} \cdot W'\Big) - k} \\ &= \frac{\beta}{\alpha} \cdot \frac{k}{\tau \cdot \left(G - \left(\frac{(-c+g)\cdot\alpha - c\cdot\beta}{\alpha\cdot(B'+g)}\right)^{\frac{n-1}{n}} \cdot W'\right) - k}. \end{aligned} \tag{C.36}$$

We derive the conditions for which $q_i, p \in [0, 1]$. For q_i we obtain

$$q_i \geq 0 \Leftrightarrow B' \geq \frac{-c \cdot (\alpha + \beta)}{\alpha} \tag{C.37}$$

and

$$q \leq 1 \Leftrightarrow (1 - \tau)^n \leq \frac{(-c + g) \cdot \alpha - c \cdot \beta}{\alpha \cdot (B' + g)} \Leftrightarrow \tau \geq 1 - \sqrt[n]{\frac{(-c + g) \cdot \alpha - c \cdot \beta}{\alpha \cdot (B' + g)}}. \tag{C.38}$$

We furthermore have

$$p \geq 0 \Leftrightarrow G \geq W' \cdot \left(\frac{(-c + g) \cdot \alpha - c \cdot \beta}{\alpha \cdot (B' + g)}\right)^{\frac{n-1}{n}} + \frac{k}{\tau} \tag{C.39}$$

and

$$p \leq 1 \Leftrightarrow G \geq W' \cdot (1 - q_i \cdot \tau)^{n-1} + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}. \tag{C.40}$$

Plugging in (C.34) into (C.40) yields

$$G \geq W' \cdot \left(\frac{(-c + g) \cdot \alpha - c \cdot \beta}{\alpha \cdot (B' + g)}\right)^{\frac{n-1}{n}} + \frac{k}{\left(1 - \frac{\beta}{\beta+\alpha}\right) \cdot \tau}. \tag{C.41}$$

Inequality (C.39) is less restrictive than inequality (C.41) and therefore not binding. Referring to condition (C.29), which is true by assumption, it is easy to see that condition (C.41) holds whenever inequalities (C.28) and (C.38) are satisfied. Together with condition (C.37), they form the binding conditions for the equilibrium to exist. One can readily see that parameter realizations which meet all conditions do exist.

Like in the deception game, the parameter sets for each equilibrium constitute a partition of the whole parameter space and are mutually exclusive. □

Proof of Proposition 3.4

According to the proof of Proposition 3.2, for $G \geq W' + \frac{k}{1-\beta}$, there exists a semi-separating equilibrium in which the identical readers check publications with positive probability. The volume of fraud is $p = \frac{\beta}{1-\beta} \cdot \frac{k}{G - W' \cdot \sqrt{\frac{g}{B'+g}} - k}$ and the overall probability of fraud detection (if published) is $1 - \frac{g}{B'+g}$.
We compare this equilibrium to two equilibria that can occur when readers are heterogeneous. Both of these equilibria are also semi-separating. In the first equilibrium, only reader 1 checks a publication with positive probability while the second reader never checks a publication. The following three conditions must hold for the equilibrium to exist:

$$\begin{aligned} \mu \cdot W + (1-\mu) \cdot G - k &= \mu \cdot W + (1-\mu) \cdot \big(W' + \Delta\big) \\ \Leftrightarrow p &= \frac{\beta}{1-\beta} \cdot \frac{k}{G - (W' + \Delta) - k}, \end{aligned} \tag{C.42}$$

$$\begin{aligned} \mu \cdot W + (1-\mu) \cdot G - k < \mu \cdot W + (1-\mu) \cdot (1-q_1) \cdot \big(W' - \Delta\big) \Leftrightarrow \\ G - \big(W' - \Delta\big) \cdot (1-q_1) < \frac{k \cdot (p \cdot (1-\beta) + \beta)}{p \cdot (1-\beta)} \end{aligned} \tag{C.43}$$

and

$$q_1 \cdot (-g) + (1-q_1) \cdot B' = 0 \Leftrightarrow q_1 = 1 - \frac{g}{B'+g}. \tag{C.44}$$

The critical conditions for obtaining $q_1, p \in [0,1]$ are $B' > 0$ and $G \geq W' + \Delta + \frac{k}{1-\beta}$. While the former condition is always true, the latter is only true whenever G is sufficiently large. By plugging in p and q_1 into (C.43), we obtain

$$\frac{g}{B'+g} < \frac{W' + \Delta}{W' - \Delta} \Leftrightarrow \Delta < W' \cdot \left(\frac{-B'}{B' + 2 \cdot g} \right). \tag{C.45}$$

Since all critical conditions can be simultaneously satisfied, the equilibrium must exist.
Facing a heterogeneous audience, the author cheats with a higher likelihood compared to a homogeneous audience if the following condition is true:

$$\begin{aligned} \frac{\beta}{1-\beta} \cdot \frac{k}{G - W' \cdot \sqrt{\frac{g}{B'+g}} - k} < \frac{\beta}{1-\beta} \cdot \frac{k}{G - (W' + \Delta) - k} \\ \Leftrightarrow \Delta > W' \cdot \left(\sqrt{\frac{g}{B'+g}} - 1 \right). \end{aligned} \tag{C.46}$$

Conditions (C.45) and (C.46) are simultaneously satisfied if

$$W' \cdot \left(\frac{-B'}{B' + 2 \cdot g}\right) > W' \cdot \left(\sqrt{\frac{g}{B' + g}} - 1\right) \Leftrightarrow \sqrt{\frac{g}{B' + g}} > \frac{2 \cdot g}{B' + 2 \cdot g} \Leftrightarrow (B')^2 > 0. \tag{C.47}$$

This condition is always true. Therefore, the author is more likely to cheat in the asymmetric equilibrium. The overall probability of fraud detection, given that a fraudulent article has been published, is identical for both kinds of audience if

$$\begin{aligned} 1 - \left(1 - \left(1 - \sqrt{\frac{g}{B' + g}}\right)\right) \cdot \left(1 - \left(1 - \sqrt{\frac{g}{B' + g}}\right)\right) = 1 - \frac{g}{B' + g} \\ \Leftrightarrow 1 - \frac{g}{B' + g} = 1 - \frac{g}{B' + g}. \end{aligned} \tag{C.48}$$

As a consequence, given that a fraudulent article has been published, the share of debunked articles is the same for both equilibria.

In the second equilibrium, the readers' roles are reversed, and it is the second reader who checks with positive probability while the first reader chooses to free ride. The necessary conditions for the equilibrium are

$$\begin{aligned} \mu \cdot W + (1 - \mu) \cdot G - k < \mu \cdot W + (1 - \mu) \cdot (1 - q_2) \cdot \left(W' + \Delta\right) \\ \Leftrightarrow G < \left(W' + \Delta\right) \cdot (1 - q_2) + \frac{k}{1 - \mu}, \end{aligned} \tag{C.49}$$

$$\begin{aligned} \mu \cdot W + (1 - \mu) \cdot G - k = \mu \cdot W + (1 - \mu) \cdot \left(W' - \Delta\right) \\ \Leftrightarrow p = \frac{\beta}{1 - \beta} \cdot \frac{k}{G - (W' - \Delta) - k} \end{aligned} \tag{C.50}$$

and

$$q_2 \cdot (-g) + (1 - q_2) \cdot B' = 0 \Leftrightarrow q_2 = 1 - \frac{g}{B' + g}. \tag{C.51}$$

The critical conditions for obtaining $q_2, p \in [0, 1]$ are $B' \geq 0$ and $G \geq W' - \Delta + \frac{k}{1-\beta}$. Both conditions are true by assumption. Substituting p and q_2 into (C.49) yields

$$G < \frac{g}{B' + g} \cdot (W' + \Delta) + G - W' + \Delta \Leftrightarrow 0 < W' \cdot \left(\frac{g}{B' + g} - 1\right) + \Delta \cdot \left(1 - \frac{g}{B' + g}\right). \tag{C.52}$$

Since the right-hand side contains only positive terms, the condition must be fulfilled and the equilibrium does exist. The volume of cheating in this equilibrium is lower than in

the symmetric equilibrium with homogeneous readers if

$$\begin{aligned} \frac{\beta}{1-\beta} \cdot \frac{k}{G - W' \cdot \sqrt{\frac{g}{B'+g}} - k} &> \frac{\beta}{1-\beta} \cdot \frac{k}{G - (W' - \Delta) - k} \\ &\Leftrightarrow \Delta > W' \cdot \left(1 - \sqrt{\frac{g}{B'+g}}\right). \end{aligned} \tag{C.53}$$

It is easy to see that this condition is always true. Since, for a cheating agent, the overall probability of getting caught is $1 - \frac{g}{B'+g}$ in both equilibria, the level of observed fraud will be lower in the asymmetric equilibrium. □

Proof of Proposition 3.5

If A produces a failure, he will prefer to publish a fraudulent article if

$$\begin{aligned} \left(1 - \left(1 - q_i^Y \cdot \tau\right)^{n^Y}\right) \cdot \left(-g^Y\right) + \left(1 - q_i^Y \cdot \tau\right)^{n^Y} \cdot B'^Y > 0 \\ \Leftrightarrow \left(1 - q_i^Y \cdot \tau\right)^{n^Y} > \frac{g^Y}{B'^Y + g^Y} \end{aligned} \tag{C.54}$$

and q_i^Y is the individual probability of checking a publication of type Y. This condition is always satisfied, due to condition (3.7).
Moreover, A will prefer mild cheating ($pub\ \widetilde{S}$) over heavy cheating ($pub\ \widetilde{L}$) if

$$\begin{aligned} &\left(1 - \left(1 - q_i^S \cdot \tau\right)^{n^S}\right) \cdot \left(-g^S\right) + \left(1 - q_i^S \cdot \tau\right)^{n^S} \cdot B'^S \geq \\ &\left(1 - \left(1 - q_i^L \cdot \tau\right)^{n^L}\right) \cdot \left(-g^L\right) + \left(1 - q_i^L \cdot \tau\right)^{n^L} \cdot B'^L \\ \Leftrightarrow \left(1 - q_i^S \cdot \tau\right)^{n^S} &\cdot \left(B'^S - g^S\right) - g^S \geq \left(1 - q_i^L \cdot \tau\right)^{n^L} \cdot \left(B'^L - g^L\right) - g^L. \end{aligned} \tag{C.55}$$

Otherwise, he will prefer heavy cheating over mild cheating.
Any R will be indifferent between checking and not checking an article of type Y if

$$\begin{aligned} \mu^Y \cdot W^Y + \left(1 - \mu^Y\right) \cdot \left(\tau \cdot G^Y + (1-\tau) \cdot \left(1 - q_i^Y \cdot \tau\right)^{n^Y - 1} \cdot W'^Y\right) - k = \\ \mu^Y \cdot W^Y + \left(1 - \mu^Y\right) \cdot \left(1 - q_i^Y \cdot \tau\right)^{n^Y - 1} \cdot W'^Y \end{aligned} \tag{C.56}$$

and $\mu_L = \frac{\beta^L}{\beta^L + (1-\beta^L-\beta^S) \cdot p^{\widetilde{L}}}$ and $\mu_S = \frac{\beta^S}{\beta^S + (1-\beta^S-\beta^L) \cdot p^{\widetilde{S}}}$.
The individidual probabilities of checking an article are therefore

$$q_i^L = \left(1 - \sqrt[n^L-1]{\frac{1}{W'^L} \cdot \left(G^L - \frac{k}{\left(1 - \frac{\beta^L}{1-\beta^S}\right) \cdot \tau}\right)}\right) \cdot \frac{1}{\tau} \tag{C.57}$$

and

$$q_i^S = \left(1 - \sqrt[n^S-1]{\frac{1}{W'^S} \cdot \left(G^S - \frac{k}{\left(1 - \frac{\beta^S}{1-\beta^L}\right) \cdot \tau}\right)}\right) \cdot \frac{1}{\tau}. \tag{C.58}$$

It is straightforward to see that $q_i^Y \in (0,1)$ whenever

$$G^Y > W'^Y + \frac{k^Y}{\left(1 - \frac{\beta^Y}{\beta^Y + (1-\beta^L-\beta^S)}\right) \cdot \tau} \tag{C.59}$$

together with an adjusted version of (3.3) holds true. Both conditions are satisfied by assumption.
Substituting (C.57) and (C.58) into (C.55) yields

$$\begin{aligned} &\left(\frac{1}{W'^S} \cdot \left(G^S - \frac{k}{\left(1 - \frac{\beta^S}{1-\beta^L}\right) \cdot \tau}\right)\right)^{\frac{n^S}{n^S-1}} \cdot (B'^S + g^S) - g^S \geq \\ &\left(\frac{1}{W'^L} \cdot \left(G^L - \frac{k}{\left(1 - \frac{\beta^L}{1-\beta^S}\right) \cdot \tau}\right)\right)^{\frac{n^L}{n^L-1}} \cdot (B'^L + g^L) - g^L \end{aligned} \tag{C.60}$$

where $\left(\frac{1}{W'^S} \cdot \left(G^S - \frac{k}{\left(1-\frac{\beta^S}{1-\beta^L}\right) \cdot \tau}\right)\right)^{\frac{n^S}{n^S-1}}$ and $\left(\frac{1}{W'^L} \cdot \left(G^L - \frac{k}{\left(1-\frac{\beta^L}{1-\beta^S}\right) \cdot \tau}\right)\right)^{\frac{n^L}{n^L-1}}$ are the overall probabilities of fraud detection for mild and heavy cheating respectively.
One can readily check that parameter constellations which make both mild cheating and heavy cheating an equilibrium do exist. To see this, assume for a moment that $G^L > G^S, W'^L = W'^S, B'^L = B'^S, g^L = g^S$ and $\beta^L = \beta^S$. Then, if G^L is sufficiently close to G^S, the inequality fails to hold (because $n^L > n^S$) and heavy cheating must constitute an equilibrium. In this equilibrium, the author is less likely to be detected overall than if he had cheated mildly. If we then increase g^L while leaving all other parameters unchanged, the inequality must eventually become fulfilled. Hence, mild cheating can likewise be the outcome of an equilibrium. □

Proof of Proposition 3.6

As a benchmark, we derive the unique equilibrium for the original game without an editor. R makes A indifferent between *care* and *no care*:

$$\begin{aligned}\mu_A \cdot B + (1-\mu_A)\cdot 0 - c = \mu_A \cdot B + (1-\mu_A)\cdot\big(q\cdot(-g) + (1-q)\cdot B'\big)\\ \Leftrightarrow q = 1 - \frac{(-c+g)\cdot\alpha - c\cdot\beta}{(B'+g)\cdot\alpha}.\end{aligned} \tag{C.61}$$

Likewise, A makes R indifferent between *check* and *no check*:

$$\mu_R \cdot W + (1-\mu_R)\cdot G - k = \mu_R \cdot W + (1-\mu_R)\cdot W' \Leftrightarrow p = \frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}. \tag{C.62}$$

One can easily conclude that the critical conditions for obtaining $q, p \in [0,1]$ are $B' \geq \frac{-c\cdot(\alpha+\beta)}{\alpha}$ and $G \geq W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)}$, which are true by assumption.

Next, we prove the existence of the crowding-out equilibrium: For A, it must be a dominant strategy not to invest care:

$$\begin{aligned}\mu_A\cdot(1-\mu_A)\cdot 0 - c < \mu_A\cdot B + (1-\mu_A)\cdot\big(r\cdot\big(q\cdot(-g)+(1-q)\cdot B'\big)+(1-r)\cdot(-f)\big)\\ \Leftrightarrow f < \frac{\frac{c\cdot(\alpha+\beta)}{\alpha} + r\cdot(q\cdot(-g-B')+B')}{1-r},\end{aligned} \tag{C.63}$$

and $\mu_A = \frac{\beta}{\beta+\alpha}$.

R makes E indifferent between *review* and *no review*:

$$\begin{aligned}\mu_E\cdot B_E + (1-\mu_E)\cdot 0 - k_E = \mu_E\cdot B_E + (1-\mu_E)\cdot\big(q\cdot(-g_E)+(1-q)\cdot B'_E\big)\\ \Leftrightarrow q = 1 - \frac{(-k_E+g_E)\cdot\alpha - k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha},\end{aligned} \tag{C.64}$$

and $\mu_E = \frac{\beta}{\beta+\alpha}$.

Moreover, E makes R indifferent between *check* and *no check*:

$$\mu_R \cdot W + (1-\mu_R)\cdot G - k = \mu_R \cdot W + (1-\mu_R)\cdot W' \Leftrightarrow r = \frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}, \tag{C.65}$$

and $\mu_R = \frac{\beta}{\beta+\alpha\cdot r}$.

The critical conditions for obtaining solutions for q and r that are within the closed unit interval are

$$B'_E \geq \frac{-k_E\cdot(\alpha+\beta)}{\alpha} \tag{C.66}$$

and

$$G \geq W' + \frac{k}{\left(1-\frac{\beta}{\beta+\alpha}\right)}. \tag{C.67}$$

The latter condition always holds true, while the former condition can easily be satisfied

for any level of α and β.
Substituting q and r into (C.63) yields

$$f < \frac{\frac{c\cdot(\alpha+\beta)}{\alpha} + \frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}\cdot\left(\left(1-\frac{(-k_E+g_E)\cdot\alpha-k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha}\right)\cdot(-g-B')+B'\right)}{1-\frac{\beta}{\alpha}\cdot\frac{k}{G-W'-k}}. \tag{C.68}$$

This condition can be satisfied without harming (C.66) and (C.67).
One can easily discern that the overall volume of published erroneous research is equal in both the original game and the modified game, viz. $\frac{\beta}{\beta+\alpha}\cdot\frac{k}{G-W'-k}$. The likelihood of a reader inspecting a published article is lower when an editor is present whenever

$$\begin{aligned} 1-\frac{(-k_E+g_E)\cdot\alpha-k_E\cdot\beta}{(B'_E+g_E)\cdot\alpha} &> 1-\frac{(-c+g)\cdot\alpha-c\cdot\beta}{(B'+g)\cdot\alpha} \\ \Leftrightarrow \frac{B'_E+g_E}{B'+g} &> \frac{(-k_E+g_E)\cdot\alpha-k_E\cdot\beta}{(-c+g)\cdot\alpha-c\cdot\beta} \end{aligned} \tag{C.69}$$

It is straightforward to see that this inequality can easily be fulfilled, e.g. if $g_E = g$, $k_E = c$ and $B'_E > B'$. □

References

Andreozzi, L. (2004). Rewarding policemen increases crime: Another surprising result from the inspection game. *Public Choice 121*(1/2), 69–82.

Azoulay, P., J. Furman, J. Krieger, and F. Murray (2015). Retractions. *Review of Economics and Statistics 97*(5), 1118–1136.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 533*(7604), 452–454.

Becker, G. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

Begley, C. and M. Lee (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*(7391), 531–533.

Bergemann, D. and U. Hege (1998). Venture capital financing, moral hazard, and learning. *Journal of Banking & Finance 22*(6), 703–735.

Bergemann, D. and U. Hege (2005). The financing of innovation: Learning and stopping. *The RAND Journal of Economics 36*(4), 719–752.

Bettis, R. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal 33*(1), 108–113.

Bhaskar, V. (2014). The ratchet effect: A learning perspective. *Working Paper*.

Bhattacharya, S. and D. Mookherjee (1986). Portfolio choice in research and development. *The RAND Journal of Economics 17*(4), 594–605.

Biais, B., T. Mariotti, G. Plantin, and J.-C. Rochet (2007). Dynamic security design: Convergence to continuous time and asset pricing implications. *The Review of Economic Studies 74*(2), 345–390.

Biais, B., T. Mariotti, J.-C. Rochet, and S. Villeneuve (2010). Large risks, limited liability, and dynamic moral hazard. *Econometrica 78*(1), 73–118.

Bobtcheff, C., B. J., and T. Mariotti (2017). Researcher's dilemma. *The Review of Economic Studies 84*(3), 969–1014.

Borch, K. (1962). Equilibrium in a reinsurance market. *Econometrica 30*(3), 424–444.

Brodeur, A., M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Bruns, S. et al. (2019). Errors and biases in reported significance levels: Evidence from innovation research. *Research Policy 48*(9).

Camerer, C. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Campbell, A., F. Ederer, and J. Spinnewijn (2014). Delay and deadlines: Freeriding and information revelation in partnerships. *American Economic Journal: Microeconomics 6*(2), 163–204.

Coase, R. (1937). The nature of the firm. *Economica 4*(16), 386–405.

Dasgupta, P. and P. David (1994). Toward a new economics of science. *Research Policy 23*(5), 487–521.

Dasgupta, P. and E. Maskin (1987). The simple economics of research portfolios. *Economic Journal 97*(387), 581–595.

Dasgupta, P. and J. Stiglitz (1980). Industrial structure and the nature of industrial activity. *The Economic Journal 90*(358), 266–293.

Di Tillio, A., M. Ottaviani, and P. Sørensen (2017). Persuasion bias in science: Can economics help? *The Economic Journal 127*, F266–F304.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution 29*(4), 605 – 610.

Erat, S. and V. Krishnan (2012). Managing delegated search over design spaces. *Management Science 58*(3), 606–623.

Felgenhauer, M. and E. Schulte (2014). Strategic private experimentation. *American Economic Journal: Microeconomics 6*(4), 74–105.

Fershtman, C. and A. Rubinstein (1997). A simple model of equilibrium in search procedures. *Journal of Economic Theory 72*(2), 432–441.

Frey, B. (2003). Publishing as prostitution - choosing between one's one ideas and academic success. *Public Choice 116*(1), 205–223.

Furman, J., K. Jensen, and F. Murray (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy 41*(2), 1747–1759.

Gall, T. and Z. Maniadis (2019). Evaluating solutions to the problem of false positives. *Research Policy 48*(2), 506–515.

Goldfarb, B. and A. King (2016). Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal 37*(1), 167–176.

Green, B. and C. Taylor (2016). Breakthroughs, deadlines, and self-reported progress: Contracting for multistage projects. *American Economic Review 106*(12), 3660–3699.

Grey, C. (2010). Organizing studies: Publications, politics and polemic. *Organization Studies 31*(6), 677–694.

Henry, E. and M. Ottaviani (2017). Research and the approval process: The organization of persuasion. *CEPR Discussion Paper No. DP11939*.

Herndon, T., M. Ash, and R. Pollin (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics 38*(2), 257–279.

Herweg, F. and D. Müller (2006). Performance of procrastinators: On the value of deadlines. *Theory and Decision 70*(3), 329–366.

Hoffmann, F. and S. Pfeil (2010). Reward for luck in a dynamic agency model. *Review of Financial Studies 23*(9), 3329–3345.

Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics 10*(2), 324–340.

Holmström, B. (1982). Moral hazard in teams. *Bell Journal of Economics 13*(1), 74–91.

Holmström, B. and P. Milgrom (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica 55*(2), 303–328.

Hörner, J. and L. Samuelson (2013). Incentives for experimenting agents. *RAND Journal of Economics 44*(4), 632–663.

Hwang, W. et al. (2005). Patient-specific embryonic stem cells derived from human SCNT blastocysts. *Science 308*(5729), 1777–1783.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine 2*(8), e124.

Ioannidis, J. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science 7*(6), 645–654.

Itoh, H. (1991). Incentives to help in multi-agent situations. *Econometrica 59*(3), 611–636.

Katolnik, S. and J. Schöndube (2019). Don't kill the goose that lays the golden eggs: Strategic delay in project completion. *Working Paper*.

Kieser, A. (2010). Unternehmen Wissenschaft. *Leviathan 38*(3), 347–367.

Kiri, B., N. Lacetera, and L. Zirulia (2018). Above a swamp: A theory of high-quality scientific production. *Research Policy 47*(5), 827–839.

Konrad, K. (2014). Search duplication in research and design spaces - Exploring the role of local competition. *International Journal of Industrial Organization 37*, 222–228.

Lacetera, N. and L. Zirulia (2011). The economics of scientific misconduct. *The Journal of Law, Economics, & Organization 27*(1), 568–603.

LaFollette, M. (1992). *Stealing into Print: Fraud, Plagarism, and Misconduct in Scientific Publishing*. University of California Press.

Lambert, R. (1983). Long-term contracts and moral hazard. *Bell Journal of Economics 14*(2), 441–452.

Legros, P. and H. Matthews (1993). Efficient and nearly efficient partnerships. *Review of Economic Studies 60*(3), 599–611.

Lewis, T. (2012). A theory of delegated search for the best alternative. *The RAND Journal of Economics 43*(3), 391–416.

Lewis, T. and M. Ottaviani (2008). Search agency. *Working Paper*.

Loury, G. (1979). Market structure and innovation. *The Quarterly Journal of Economics 93*(3), 395–410.

Malcomson, J. and F. Spinnewyn (1988). The multiperiod principal-agent problem. *The Review of Economic Studies 55*(3), 391–407.

Manso, G. (2011). Motivating innovation. *Journal of Finance 66*(5), 1823–1860.

Mason, R. and J. Välimäki (2015). Getting it done: Dynamic incentives to complete a project. *Journal of the European Economic Association 13*(1), 62–97.

McElreath, R. and P. Smaldino (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS One 10*(8), e0136088.

Merton, R. (1963). Resistance to the systematic study of multiple discoveries in science. *European Journal of Sociology 4*(2), 239–249.

Merton, R. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press.

Mookherjee, D. (1984). Optimal incentive schemes with many agents. *Review of Economic Studies 51*(3), 433–446.

Necker, S. (2014). Scientific misbehavior in economics. *Research Policy 43*(10), 1747–1759.

Nissen, S., T. Magidson, K. Gross, and C. Bergstrom (2016). Publication bias and the canonization of false facts. *eLife 5*, e21451.

O'Donoghue, T. and M. Rabin (1999). Incentives for procrastinators. *The Quarterly Journal of Economics 114*(3), 769–816.

Page, S. (2006). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Parkinson, C. (1957). *Parkinson's Law, and Other Studies in Administration*. Random House.

Planck, M. (1932). Epilogue: A socratic dialogue. Planck - Einstein - Murphy. In *Where is Science going?* Allen & Ulwin/W.W. Norton.

Reinhart, C. and K. Rogoff (2010). Growth in a time of debt. *American Economic Review: Papers & Proceedings 100*(2), 573–578.

Sannikov, Y. (2008). A continuous-time version of the principal-agent problem. *The Review of Economic Studies 75*(3), 957–984.

Simmons, J., L. Nelson, and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*(11), 1359–1366.

Starbuck, W. (2005). How much better are the most prestigious journals? The statistics of academic publication. *Organization Science 16*(2), 180–200.

Stephan, P. (1996). The economics of science. *Journal of Economic Literature 34*(3), 1199–1235.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter than the Few*. Doubleday.

Townsend, R. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory 21*(2), 265–293.

Toxvaerd, F. (2006). Time of the essence. *Journal of Economic Theory 129*(1), 252–272.

Tsebelis, G. (1989). The abuse of probability in political analysis: The Robinson Crusoe fallacy. *American Political Science Review 83*(1), 77–91.

Tsebelis, G. (1990). Penalty has no impact on crime: A game-theoretic analysis. *Rationality and Society 2*(3), 255–286.

Ulbricht, R. (2016). Optimal delegated search with adverse selection and moral hazard. *Theoretical Economics 11*(1), 253–278.

Wagenmakers, E., R. Wetzels, D. Borsboom, and H. van der Maas (2011). Why psychologists must change the way they analyze their data: The case of PSI: Comment on BEM (2011). *Journal of Personality and Social Psychology 100*(3), 426 – 432.

Weber, M. (1946). Science as a vocation. In *From Max Weber: Essays in sociology*, pp. 129–156. Oxford University Press.

Wible, J. (1998). *The Economics of Science: Methodology and Epistemology as if Economics Really Mattered*. Routledge.

Williams, N. (2010). A solvable continuous time dynamic principal-agent model. *Journal of Economic Theory 159*(B), 989–1015.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende schriftliche Arbeit selbstständig und ohne fremde Hilfe verfasst, eventuelle Beiträge von Koautoren dokumentiert, keine anderen als die von mir angegebenen Hilfsmittel benutzt und alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet, sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht habe.

Matthias Verbeck Frankfurt am Main, den 6. Dezember 2021