Essays on the Foundation of Beliefs and Preferences in Economics

Simon Gleyze

Université Paris 1 Panthéon - Sorbonne
UFR02 Paris Jourdan Sciences Économiques
Paris School of Economics

THÈSE

pour l'obtention du titre de Docteur en Sciences Économiques
Présentée et soutenue publiquement le 3 Juin 2022 par:

Simon Gleyze

Essays on the Foundation of Beliefs and Preferences in Economics

Sous la direction de:
Philippe Jehiel, Professeur à PSE & UCL
Jean-Marc Tallon, Professeur à PSE & Paris 1 Panthéon-Sorbonne

Membres du Jury:

Renato Gomez	Rapporteur	Professeur à TSE Dir. de Recherche au CNRS
Ronny Razin	Rapporteur	Professeur à LSE
Olivier Tercieux	Examinateur	Professeur à PSE Dir. de Recherche au CNRS
Jean Tirole	Examinateur	Professeur à TSE

Acknowledgments

I owe to the people who guided me in my academic journey. First, my advisors Philippe Jehiel and Jean-Marc Tallon. They taught me the fine art of recognizing good research, and showed me the challenges one faces when trying to produce quality work. My journey also greatly benefited from many discussions with Olivier Tercieux who always treated me as an equal.

I also want to thank particularly my coauthors, for putting up with my errors and nonsensical ideas. This dissertation is truly a collaborative work, and many more ideas than I would like to admit are theirs. My collaboration with Agathe Pernoud is both my first and my most productive one. Many ideas that I have in economics and about economics were the result of our discussions. My collaboration with Niels Boissonnet and Alexis Ghersengorin was the most chaotic but also deepest one. I greatly enjoyed our discussions and I am very proud of our work on preference change. My collaboration with Philippe Jehiel was very important at a personal level because the under-representation of disadvantaged students at elite colleges is something I personally care about.

Finally, I am thankful to my family and especially my parents for putting up with my PhD during five years even though they have no idea what I have been doing. My brother, Valentin Gleyze, for paving the way of my academic journey until the PhD. All the friends I made along the way at Bordeaux and Paris during my studies. Most importantly, Agathe Pernoud for being my partner in crime during all these years, and for bringing me so much joy during this tough experience. I owe her a lot for where I am today, and this thesis is dedicated to her.

Summary

Beliefs and preferences are two building blocks which, together with the concept of rationality, underlie most descriptive and normative analysis in economics. This dissertation explores the (old) idea that beliefs and preferences are not fixed, but instead vary in reaction to the environment. Chapter 1, joint with Agathe Pernoud, studies informational incentives in mechanism design, that is how the institutional environment affects how agents acquire information. Agents have uncertainty on their preferences and can buy information on their valuations as well as others'. We show that, under certain conditions, agents acquire information on others' whenever the mechanism violates a separability condition which rules out most economically meaningful mechanisms. Chapter 2, joint with Philippe Jehiel, studies how students form their subjective admission chances. We propose a model in which students average the past experience of their peers which induces a regression to the mean of admission chances. Two inefficiencies arise: high-achieving disadvantaged students self-select out of elite colleges, and low-achieving advanteged students wastefully apply to elite colleges even though their admission chances are zero. Chapter 3, joint with Niels Boissonnet and Alexis Ghersengorin, investigates the question of preference change. We show that it is possible to falsify and identify a model of preference change in which the decision maker changes her behavior according to a meta-preference relation. Chapter 4, joint with Agathe Pernoud, studies how beliefs about economic models affect strategic communication, for instance between voters and medias. We show that holding a misspecified model can increase the informativeness of communication, and model misspecification is stable as communication on models by a third party is generally limited.

Résumé

Les croyances et les préférences, ainsi que le concept de rationalité, sont les fondements qui sous-tendent l'analyse descriptive et normative de l'économie. Cette thèse explore l'idée (ancienne) que les croyances et les préférences ne sont pas fixes, mais changent face à l'environement. Le premier chapitre, coécrit avec Agathe Pernoud, étudie les incitations informationalles au sein de la théorie des incitations. On suppose que les agents sont incertains sur leurs préférences et peuvent acheter de l'information dessus ainsi que sur celles des autres agents. On montre que, sous certaines conditions, les agents acquièrent de l'information sur les préférences des autres dès que le mécanisme viole une condition de séparabilité qui exclue la majorité des mécanismes qui représentent un intérêt économique. Le deuxième chapitre, coécrit avec Philippe Jehiel, étudie la manière dont les étudiants forment leurs croyances sur leurs chances d'admission dans l'enseignement supérieur. On propose un modèle dans lequel les étudiants utilisent la moyenne des admissions passées de leurs pairs, ce qui entraine une regréssion vers la moyenne de leurs croyances. Ceci entraine deux anomalies: les étudiants doués mais désavantagés s'auto-selectionnent, et les étudiants moyens mais avantagés candidatent aux formations sélectives alors que leur probabilité d'admission est nulle. Le troisième chapitre, coécrit avec Niels Boissonnet et Alexis Ghersengorin, explore la question du changement de préférence. On montre qu'il est possible de falsifier et d'identifier un modèle de changement de préférence dans lequel l'agent modifie sont comportement en suivant une meta-préférence. Le quatrième chapitre, coécrit avec Agathe Pernoud, étudie comment les croyances sur les modèles économiques affectent la communication stratégique, par exemple entre les voters et les médias. On montre que détenir un modèle mal spécifié peut augmenter l'informativité de la communication, et que ces erreurs de spécification sont stables puisque la communication sur les modèles par une tierce personne est limitée.

Contents

Introduction

Agrippa's trilemma is an argument reported by Sextus Empiricus intended to demonstrate the impossibility of proving any truth—or at least, without accepting certain assumptions. The trilemma asserts that there are only three ways to complete a proof: the circular argument, in which the proof of some proposition presupposes the truth of that very proposition; the regressive argument, in which each proof must rest on an additional proof, ad infinitum; and the dogmatic argument, which rests on precepts that are simply accepted as true.

Here, I will argue that our approach to preferences and beliefs in economics follows a dogmatic tradition (in the aforementioned sense) that rests on normative principles. Then, I will advocate for regressive or circular foundations of preferences and beliefs, discussing their advantages and limitations, and how my dissertation contributes to this approach.

The dogmatic tradition in economics dates back to at least the *Theory of Games and Economic Behavior* by von Neumann and Morgenstern published in 1947. They introduce four axioms of "rationality" on agents' preferences which guarantee that their behavior can be represented as if they were maximizing a utility function. The axioms are justified on normative grounds: any reasonable, internally consistent individual should readily accept them. This agenda is further developed by Savage (1954) who provides a foundation of beliefs solely based on preferences. Again, the idea is that if one accepts a set of reasonable precepts, then agents' behavior can be described as if they hold subjective beliefs and maximize subjective expected utility.

This dogmatic approach is oblivious to the difficult question of the *origin* of preferences and beliefs. Instead, if accepted, these axioms simply guarantee the *existence* of beliefs and rational preferences. This, however, does not imply that preferences and beliefs are fixed, i.e. not responsive to the environment. Many (of course not all) economists, however, seem to commit this logical fallacy. I believe this is the case for at least two reasons: First, the dogmatic approach, by its axiomatic nature, is not intended as a reflection on the origin of

preferences and beliefs. Second, indexing the preference ordering $\succsim$ (or prior belief μ_0) by the environment $\succsim^{\mathcal{E}}$ (or $\mu_0^{\mathcal{E}}$) is useless unless one is willing to make further restrictions on how the reasonableness of each axiom depends on the environment—which seems a quite difficult and controversial endeavor.

The circular approach is very common in economics following the work of Nash in game theory, and the work of Arrow, Debreu, and McKenzie on general equilibrium theory in the 1950s. The circularity is used to explain equilibrium behavior in strategic contexts, or equilibrium prices, but more rarely preferences and beliefs. Instead, the latter are typically fixed and taken as primitives in the analysis. There are two important exceptions though: the literature on herd behavior where beliefs are derived by observing other agents' behavior, and the literature on strategic communication (with or without commitment).

The regressive approach is more rare, the most notable examples being the concepts of common knowledge (Aumann, 1976) and rationalizability (Bernheim, 1984) in game theory, or—to some extent—the ideas of second order acts in the smooth ambiguity model[1] (Klibanoff, Marinacci and Mukerji, 2005) and the level-k approach. There is always some discomfort in the regressive argument because, either it is conducted ad infinitum and it is unclear how we should get there, or the regression is truncated and the stopping point will appear arbitrary to some.

While the circular and regressive approaches are rarely used as foundations for beliefs and preferences, they have a major advantage compared to the axiomatic approach. Indeed, the former put tight constraints—usually optimization or consistency constraints—on how preferences or beliefs evolve with the environment. Instead, the axiomatic approach is much more flexible and puts a priori no such constraints. This added flexibility makes it harder to come up with convincing axioms that are environment-specific.

In Chapter 1 of my dissertation, joint with Agathe Pernoud, we take a circular approach to the formation of beliefs in mechanism design. Instead of

[1]Note that this paper combines both the axiomatic and regressive approaches. Nevertheless, in light of the controversy created by second-order acts, I will classify it in the regressive approach.

assuming that agents' private information is fixed, we endow agents with a technology of costly information acquisition which allow them to optimize the granularity of their private information depending on the environment (i.e., other players' preferences, mechanism chosen by the designer, etc.). Note that there is still a dependence to the prior belief which is "outside the model": information acquisition will not be the same depending on which states are more likely ex-ante. Though, agents' private information after the information acquisition stage is clearly endogenous to the environment.

More specifically, we are interested in whether agents acquire information on other's preferences. We call a mechanism *informationally simple* if there is an equilibrium in which agents do not acquire information on others. This is interesting for two reasons: First, informational simplicity is a necessary condition for the existence of dominant strategy to the overall game that includes information acquisition. Second, informational simplicity guarantees independent and private values (IPV) at the interim stage, which has been the standard assumption in mechanism design. Therefore, we are able to assess whether this assumption is likely to arise endogenously. We show that, under a smoothness and Inada condition on information acquisition costs, only mechanisms that satisfy a separability property are informationally simple. This separability property, however, rules out most economically meaningful mechanisms. Therefore, for most mechanisms used in practice such as VCG, agents will acquire information on others.

In Chapter 2, joint with Philippe Jehiel, we again take a circular approach to the formation of beliefs. Instead of considering information acquisition, we take a non-Bayesian approach based on non-parametric estimation. We assume that students use past admissions of their peers to estimate the distribution of admission chances conditional on application. We require that actions are rational given subjective beliefs, and subjective beliefs are consistent with respect to the action profile. Hence, beliefs are endogenous to the environment (i.e., other students preferences, capacities at colleges, etc.) but very much constrained by the fact that we impose a specific estimator that must be consistent with equilibrium play.

We first look at the one-neighborhood case and show that two types of in-

efficiencies arise: First, high-achieving disadvantaged students self-select out of elite colleges. Second, average ability advantaged students wastefully apply to elite colleges even though their true admission chances are zero. These inefficiencies arise because the non-parametric estimation induces a reversion to the mean of subjective admission chances. We then investigate how competing neighborhoods affect welfare and the quality of admitted students, as well as the efficacy of various policy instruments such as quotas or mixed neighborhoods.

In Chapter 3, joint with Niels Boissonnet and Alexis Ghersengorin, we take a regressive approach (together with an axiomatic approach) to the foundation of preferences. We are interested in the (old) question: where do preferences come from? Instead of assuming that we are born with fixed preferences—which contradicts both common intuition and empirical evidence—we assume that agents can reevaluate their preferences when they become aware of new attributes of the alternatives. For instance, it has been shown in political economy that nationalist discourses on immigration change voters' beliefs but also voting intention, even when beliefs have been recalibrated following fact checking. This suggests that politicians, by raising awareness on identity issues, may actually change people's preferences.

We show that if individuals change preferences consistently, meaning that when they are aware twice of the same set of attributes they do not revert to their previously hold preferences, then we can represent their behavior as if they were maximizing a meta-preference relation. This can capture individuals who try to align their choice behavior with their values. One may ask: where do such values come from then? Why should we stop at the second order, and not continue ad infinitum? This is a valid objection, but in practice it could be that most of the explanatory power is derived from the first few orders. Hence, a truncated regressive argument can have value even though it is not fully "solving" the question of where preferences come from.

In Chapter 4, joint with Agathe Pernoud, we take a circular approach to the formation of beliefs in a cheap talk game. In particular, we are interested in the impact of holding misspecified beliefs on economic models on equilibrium communication. We introduce a framework in which Receiver communicates

on a multi-dimensional state with a partially misaligned Sender, but not all dimensions are relevant. A subjective model describes which dimensions Receiver believes are relevant.

We show that holding a simple, possibly misspecified, model can increase communication between Sender and Receiver. The intuition is that holding a simple model acts as a commitment device on individually rational actions that limit the scope of information manipulation by Sender. We then investigate whether communication on models is feasible. We introduce a Principal who is informed about the true model and who is perfectly aligned with Receiver. Despite preference alignment, we show that communication on models is impossible due to the instrument value of misspecification when communicating on states in the second stage.

Overall, this suggests that foundations of preferences and beliefs that rely on circular or regressive arguments provide natural constraints with the economic environment that would otherwise be difficult to justify using an axiomatic approach. This dissertation contributes to this research agenda which deserves, I believe, further theoretical and empirical investigation in the future.

Introduction

Le trilemme d'Agrippa est un argument rapporté par Sextus Empiricus visant à démontrer l'impossibilité de prouver toute vérité — ou du moins, sans accepter certaines hypothèses. Le trilemme affirme qu'il n'y a que trois façons de compléter une preuve : l'argument circulaire, dans lequel la preuve d'une proposition présuppose la vérité de cette même proposition ; l'argument régressif, dans lequel chaque preuve doit reposer sur une preuve supplémentaire, à l'infini ; et l'argument dogmatique, qui repose sur des préceptes qui sont simplement acceptés comme vrais.

Je soutiendrai ici que notre approche des préférences et des croyances en économie s'inscrit dans une tradition dogmatique (au sens susmentionné) qui repose sur des principes normatifs. Ensuite, je plaiderai pour des fondements régressifs ou circulaires des préférences et des croyances, en discutant de leurs avantages et de leurs limites, et en expliquant comment ma thèse contribue à cette approche.

La tradition dogmatique en économie remonte a minima à l'ouvrage *Theory of Games and Economic Behavior* de von Neumann et Morgenstern publié en 1947. Ils introduisent quatre axiomes de "rationalité" sur les préférences des agents qui garantissent que leur comportement peut être représenté comme s'ils maximisaient une fonction d'utilité. Ces axiomes sont justifiés par des raisons normatives : tout individu raisonnable et cohérent devrait les accepter sans hésiter. Ce programme est approfondi par Savage (1954) qui propose un fondement des croyances uniquement basé sur les préférences. Là encore, l'idée est que si l'on accepte un ensemble de préceptes raisonnables, alors le comportement des agents peut être décrit comme s'ils détenaient des croyances subjectives et maximisaient l'utilité espérée subjective.

Cette approche dogmatique ne tient pas compte de la question difficile de l'origine des préférences et des croyances. Au lieu de cela, s'ils sont acceptés, ces axiomes garantissent simplement l'*existence* des croyances et des préférences rationnelles. Cela n'implique toutefois pas que les préférences et les croyances

sont fixes, c'est-à-dire qu'elles ne sont pas dépendantes de l'environnement. De nombreux économistes (pas tous, bien sûr) semblent toutefois commettre ce sophisme logique. Je pense que c'est le cas pour au moins deux raisons : premièrement, l'approche dogmatique, de par sa nature axiomatique, n'est pas conçue comme une réflexion sur l'origine des préférences et des croyances. Deuxièmement, la simple indexation des préférences $\succsim$ (ou de la croyance initiale μ_0) par l'environnement $\succsim^{\mathcal{E}}$ (ou $\mu_0^{\mathcal{E}}$) est inutile à moins que l'on soit prêt à faire des restrictions supplémentaires sur la façon dont la vraisemblance de chaque axiome dépend de l'environnement — ce qui semble être une entreprise difficile et controversée.

L'approche circulaire est très courante en économie suite aux travaux de Nash en théorie des jeux, et aux travaux d'Arrow, Debreu, et McKenzie sur la théorie de l'équilibre général dans les années 1950. La circularité est utilisée pour expliquer le comportement d'équilibre dans des contextes stratégiques, ou les prix d'équilibre, mais plus rarement les préférences et les croyances. Au contraire, ces dernières sont généralement fixées et prises comme primitives dans l'analyse. Il y a cependant deux exceptions notables : la littérature sur le *herding* où les croyances sont dérivées de l'observation du comportement des autres agents, et la littérature sur la communication stratégique (avec ou sans engagement).

L'approche régressive est plus rare, les exemples les plus notables étant les concepts de connaissance commune (Aumann, 1976) et de rationalisabilité (Bernheim, 1984) dans la théorie des jeux, ou — dans une certaine mesure — les idées d'actes de second ordre dans le modèle d'ambiguïté continue (Klibanoff, Marinacci et Mukerji, 2005) et l'approche de niveau-k. L'argument régressif suscite toujours un certain malaise car, soit il est mené à l'infini et on ne sait pas toujours comment y parvenir, soit la régression est tronquée et le point d'arrêt paraîtra arbitraire à certains.

Si les approches circulaire et régressive sont rarement utilisées comme fondements des croyances et des préférences, elles présentent un avantage majeur par rapport à l'approche axiomatique. En effet, la première impose des contraintes strictes — généralement des contraintes d'optimisation ou de consistence — sur la manière dont les préférences ou les croyances évoluent avec

l'environnement. Au contraire, l'approche axiomatique est beaucoup plus flexible et n'impose a priori aucune contrainte de ce type. Cette flexibilité supplémentaire rend plus difficile l'élaboration d'axiomes convaincants qui soient spécifiques à l'environnement.

Dans le chapitre 1 de ma thèse, en collaboration avec Agathe Pernoud, nous adoptons une approche circulaire de la formation des croyances dans le contexte de la théorie des mécanismes optimaux. Au lieu de supposer que l'information privée des agents est fixe, nous dotons les agents d'une technologie d'acquisition d'information coûteuse qui leur permet d'optimiser la granularité de leur information privée en fonction de l'environnement (i.e. les préférences des autres joueurs, le mécanisme choisi par l'autorité publique, etc.). Notez qu'il y a toujours une dépendance à la croyance initiale qui est "hors du modèle" : l'acquisition d'information ne sera pas la même selon quelles états sont les plus probables ex-ante. Cependant, l'information privée des agents après l'étape d'acquisition de l'information est bien endogène à l'environnement.

Plus précisément, nous nous intéressons à la question de savoir si les agents acquièrent des informations sur les préférences des autres. Nous appelons un mécanisme *informationnellement simple* s'il existe un équilibre dans lequel les agents n'acquièrent pas d'informations sur les autres. Ceci est intéressant pour deux raisons : premièrement, la simplicité informationnelle est une condition nécessaire à l'existence d'une stratégie dominante au jeu global qui inclut l'acquisition d'information. Deuxièmement, la simplicité informationnelle garantit des valeurs indépendantes et privées (IPV) à l'étape intermédiaire, ce qui est l'hypothèse standard dans la théorie des mécanismes optimaux. Par conséquent, nous sommes en mesure d'évaluer si cette hypothèse est susceptible d'apparaître de manière endogène. Nous montrons que, sous une condition de continuité et avec une hypothèse d'Inada sur les coûts d'acquisition de l'information, seuls les mécanismes qui satisfont une propriété de séparabilité sont simples sur le plan informationnel. Cette propriété de séparabilité, cependant, exclut la plupart des mécanismes économiquement pertinents. Par conséquent, pour la plupart des mécanismes utilisés dans la pratique, comme le mécanisme VCG, les agents vont acquérir des informations sur les autres.

Dans le chapitre 2, en collaboration avec Philippe Jehiel, nous adoptons

une approche circulaire de la formation des croyances. Au lieu de considérer de l'acquisition d'information, nous adoptons une approche non-Bayesienne basée sur l'estimation non-paramétrique. Nous supposons que les étudiants utilisent les admissions passées de leurs pairs pour estimer la distribution des chances d'admission conditionnelles à la candidature. Nous exigeons que les actions soient rationnelles compte tenu des croyances subjectives, et que les croyances subjectives soient cohérentes par rapport au profil d'action. Par conséquent, les croyances sont endogènes à l'environnement (c'est-à-dire les préférences des autres étudiants, les capacités des collèges, etc.) mais sont très contraintes par le fait que nous imposons un estimateur spécifique qui doit être cohérent avec les actions d'équilibre.

Nous nous penchons d'abord sur le cas d'un seul quartier et montrons que deux types d'inefficacités apparaissent : premièrement, les étudiants défavorisés très performants s'auto-sélectionnent en dehors des universités d'élite. Deuxièmement, les étudiants favorisés aux performances moyennes postulent inutilement aux universités d'élite, même si leurs chances réelles d'admission sont nulles. Ces inefficacités sont dues au fait que l'estimation non-paramétrique induit une régression à la moyenne des chances d'admission subjectives. Nous étudions ensuite comment la compétition entre quartiers affecte le bien-être et la qualité des étudiants admis, ainsi que l'efficacité de divers instruments de politique publique tels que les quotas ou les quartiers mixtes.

Dans le chapitre 3, coécrit avec Niels Boissonnet et Alexis Ghersengorin, nous adoptons une approche régressive (ainsi qu'une approche axiomatique) du fondement des préférences. Nous nous intéressons à la (vieille) question : d'où viennent les préférences ? Au lieu de supposer que nous naissons avec des préférences fixes — ce qui contredit à la fois l'intuition commune et les résultats empiriques — nous supposons que les agents peuvent réévaluer leurs préférences lorsqu'ils prennent connaissance de nouveaux attributs des alternatives. Par exemple, il a été démontré en économie politique que les discours nationalistes sur l'immigration modifient les croyances des électeurs mais aussi leur intention de vote, même lorsque les croyances ont été recalibrées par du *fact checking*. Cela suggère que les politiciens, en sensibilisant aux questions d'identité, peuvent effectivement changer les préférences des gens indépen-

damment des croyances.

Nous montrons que si les individus changent de préférences de manière cohérente, c'est-à-dire que lorsqu'ils prennent conscience deux fois du même ensemble d'attributs, ils ne reviennent pas à leurs préférences antérieures, alors nous pouvons représenter leur comportement comme s'ils maximisaient une relation de méta-préférence. Cela permet de rendre compte d'individus qui tentent d'aligner leurs choix avec leurs valeurs. On peut tout de même se demander : d'où viennent alors ces valeurs ? Pourquoi devrions-nous nous s'arrêter au deuxième ordre, et non pas continuer à l'infini ? Il s'agit d'une objection valable, mais en pratique, il se peut que la majeure partie du pouvoir explicatif provienne des premiers ordres. Par conséquent, un argument régressif tronqué peut avoir de la valeur même s'il ne "résout" pas complètement la question de l'origine des préférences.

Dans le chapitre 4, en collaboration avec Agathe Pernoud, nous adoptons une approche circulaire de la formation des croyances dans un jeu de *cheap talk*. En particulier, nous nous intéressons à l'impact des croyances mal spécifiées vis-à-vis de modèles économiques sur la communication à l'équilibre. Nous introduisons un cadre dans lequel le récepteur communique sur un état multidimensionnel — mais dont toutes les dimensions ne sont pas pertinentes — avec un émetteur dont les préférences divergent partiellement. Un modèle subjectif décrit les dimensions que le récepteur croit être pertinentes.

Nous montrons que la détention d'un modèle simple, éventuellement mal spécifié, peut augmenter la communication entre l'émetteur et le récepteur. L'intuition est que la détention d'un modèle simple agit comme un dispositif d'engagement sur des actions rationnelles qui limitent la portée de la manipulation de l'information par l'émetteur. Nous examinons ensuite si la communication sur les modèles est réalisable. Nous introduisons un Principale qui est informé du vrai modèle et qui est parfaitement aligné avec le récepteur. Malgré l'alignement des préférences, nous montrons que la communication sur les modèles est impossible en raison de la valeur instrumentale de la mauvaise spécification lors de la communication sur les états dans la deuxième étape.

Dans l'ensemble, cela suggère que les fondements des préférences et des croyances qui reposent sur des arguments circulaires ou régressifs fournissent

des contraintes naturelles avec l'environnement économique qui seraient autrement difficiles à justifier en utilisant une approche axiomatique. Cette thèse contribue à ce programme de recherche qui mérite, je crois, d'être approfondi sur le plan théorique et empirique dans le futur.

Chapter 1

Informationally Simple Incentives[1]

Scholars have long understood that institutions can have a strong impact on the formation of preferences. Surprisingly little attention, however, has been paid to how institutions shape the ways in which we constitute our knowledge and acquire information, that is, to how agents' *informational incentives* are affected by institutional rules. This can sometimes be of primary importance because heterogeneous informational incentives can lead to unequal opportunities in voting, labor market outcomes, education choices, investment decisions, etc. Conversely, little is known about how these informational incentives constrain the type of institutions that are actually implementable. In this paper, we make progress toward addressing these questions.

We investigate what kind of mechanisms lead to simple informational incentives, and why simple informational incentives matter for the design of institutions. We consider good allocation problems in which agents' valuations for the good are private, and independently drawn. Agents are uncertain about their preferences, but can acquire information about their preferences as well as others' before entering the mechanism. For instance, students facing a school

[1]This paper is joint with Agathe Pernoud. We thank Emir Kamenica and three anonymous referees for suggestions that greatly improved the paper. We are grateful to Piotr Dworczak, Fuhito Kojima, Shengwu Li, Elliot Lipnowski, Mike Ostrovsky, Eduardo Perez-Richet, Doron Ravid, Al Roth, Ilya Segal, Olivier Tercieux and especially Matt Jackson, Philippe Jehiel and Paul Milgrom for helpful conversations and comments. We also thank seminar participants at PSE, Stanford, Akbarpour–Milgrom discussion group, YES2020, and the CEME Decentralization Conference (2021) for valuable comments and questions. S. Gleyze acknowledges the support of the EUR grant ANR-17-EURE-0001.

choice mechanism can not only acquire information on their own preferences for the different schools but also learn about how demanded they are. Similarly, bidders in an auction mechanism can not only learn about their own valuation for the good, but also consult firms to gauge the toughness of the competition. Informational simplicity is defined as acquiring information on one's own preferences only, and not on others'.

One might think that strategy-proofness guarantees informational simplicity since it implies agents have a dominant strategy at the interim stage. Our main result is that this is however not the case: for a large set of information acquisition cost functions, and whenever the mechanism violates a separability condition[2]—which is the case of most economically meaningful mechanisms—, players always have an incentive to learn about others' preferences even though they are not *directly* payoff-relevant. In particular, even strategy-proof mechanisms such as VCG incentivize players to acquire information on others'. Moreover, we show that such informational incentives make strategy-proof mechanisms no longer dominant solvable at the ex-ante stage, when players decide what information to acquire. These results hold whenever the cost of information satisfies an Inada condition—which makes it never optimal to become *fully* informed about any state—and a smoothness condition—which guarantees that agents can fine-tune the informativeness of signals without discontinuously changing their cost. Importantly, the result holds even though players' underlying preferences are independent and private. Otherwise, players would have a direct incentive to learn about the preferences of others as it would be informative on their own preferences.

The intuition behind our main result is the following: The set of outcomes that a player can bring about, call this her *opportunity set*, depends on other players' reports to the designer, and hence on the entire vector of fundamentals determining the preferences of the population. Since the value of information on her own preferences depends on her opportunity set, it *indirectly* depends

[2]Say a mechanism is separable if agents' reports do not interact with one another in the allocation function: for all i, all $m_i, m_i' \in M_i$, and all $m_{-i}, m_{-i}' \in M_{-i}$, $x_i(m_i, m_{-i}) - x_i(m_i', m_{-i}) = x_i(m_i, m_{-i}') - x_i(m_i', m_{-i}')$. Among mechanisms that have received some attention in the literature, only dictatorial mechanisms satisfy such separability condition. All standard auction formats do not.

on other players' preferences as well. If gathering a little bit of information about others' preferences is not costlier than learning additional information about her own, then it is generically optimal for the player to devote resources to acquiring information about others' preferences first. That helps her predict others' report, allowing her to acquire more information on herself when it is more valuable.

Finally, we explore the implications of our result for mechanism design. First, it appears that strategic simplicity, as captured by strategy-proofness, is more limited than previously thought. Much of the literature focuses on strategic simplicity at the interim stage, that is once players have acquired their private information. Our results show that strategy-proofness does not guarantee strategic simplicity at the ex-ante stage of information acquisition: Indeed, only informationally simple mechanisms admit equilibria in dominant strategies in the extended game. This is one argument as to why informational simplicity might be valuable in practice: it ensures agents have a dominant strategy when deciding what information to acquire, leading to more robust predictions and fewer strategic mistakes. However, our main result implies that only *de facto* separable mechanisms satisfy such property.

Second, a direct corollary of our main result is that the standard Independent Private value (IPV) assumption is unlikely to arise endogenously. Of course, this assumption is usually understood as a technically convenient approximation of reality—nothing more. Nevertheless, our result makes precise why this is unlikely to hold in practice, and why departures from IPV in the standard framework lead to discontinuities such as Crémer and McLean (1988)'s full surplus extraction result. Instead, our approach restores a form of continuity: side bets at the interim stage reduce the amount of information acquisition at the ex-ante stage—in particular, side bets prevent the efficient amount of information acquisition. Hence *constrained* surplus extraction is feasible, but *full* surplus extraction is not because players internalize the informational incentives generated by side bets.

Related Literature. A first strand of the literature investigates information acquisition with fixed mechanisms. Persico (2000) proves a representation the-

orem for the demand for information in several auction formats. Bergemann et al. (2009) show that with interdependent values the equilibrium level of information acquisition is inefficient under VCG. More recently, Bobkova (2019) investigates the incentives to learn about private versus common value components in auctions.

A second strand of the literature investigates optimal mechanism design with information acquisition. Bergemann and Välimäki (2002) show that in a standard allocation problem with monetary transfers, private values, and information acquisition on own preferences only, VCG is ex-ante efficient. Hatfield et al. (2018) strengthen this result by showing that strategy-proofness is also necessary for ex-post efficient mechanisms to induce ex-ante efficient information acquisition. Interestingly, a corollary of our result is that VCG induces ex-ante *inefficient* information acquisition when agents are allowed to learn about others' preferences in addition to their own, as it endogenously leads to interdependent values at the interim stage.

In school choice settings, Immorlica et al. (2018) look for mechanisms that are stable and induce students to acquire information efficiently. Roesler and Szentes (2017) and Ravid et al. (2019) consider monopoly pricing when buyers can flexibly acquire information. In their papers, the seller chooses the mechanism after the buyer chooses her information strategy, whereas in our paper the designer ex-ante commits to a mechanism. Hence in their model the buyer must internalize the seller's strategy, whereas in our paper the designer must internalize the agents' future decisions (what we refer to as "informational incentives"). Mensch (2019) considers a screening problem with informational incentives and characterizes the optimal mechanism.

Most of the literature investigates information acquisition on one's own preferences only.[3] In this paper, we allow agents to acquire information on others as well, and investigate when it would be optimal for them to do so.

[3] A notable exception is Larson and Sandholm (2001) in the computer science literature. They introduce a model in which agents can devote computational resources to discover their own as well as others' valuation. For several auction formats, they show that players compute the valuation of others in equilibrium.

1.1 Motivating Example

Two bidders compete in a Second-Price auction to acquire a good. Contrary to the standard approach, bidders are uncertain about their valuation for the good. Without uncertainty on their valuation, bidders would simply play their dominant strategy which is to bid their true valuation. Bidder 1's valuation is either high $\overline{\omega}_1 \in \Omega_1$ or low $\underline{\omega}_1 \in \Omega_1$, and similarly for bidder 2. We denote the state space by $\Omega = \Omega_1 \times \Omega_2$, and suppose agents' valuations are independently drawn from a uniform distribution, so all four states are equally likely. Let $\overline{\omega}_1 > \overline{\omega}_2 > \underline{\omega}_1 > \underline{\omega}_2$, such that if bidders knew their own valuation for the good and played their dominant strategy, bidder 2 would only win the auction in state $(\underline{\omega}_1, \overline{\omega}_2)$.

Bidders can engage in costly information acquisition about $\omega = (\omega_1, \omega_2)$ ex ante. Importantly, they can privately acquire information on *any* fundamental, and hence not only learn about their own valuation but also about the other bidder's if they wish to. Information is costly, and bidders trade-off the value and cost of information upon acquiring it. For this example only, we consider the entropic cost function, which has been introduced in the rational inattention literature. Informally, the cost of information is proportional to the expected reduction in uncertainty as measured by the entropy of beliefs:

$$\text{cost of information} = \lambda\Big(\text{prior entropy} - \mathbb{E}[\text{posterior entropy}]\Big)$$

where λ is a scaling parameter. This cost function satisfies the key assumptions we impose for our main result: it is smooth, and the marginal cost of becoming fully informed is unbounded (Inada condition). We discuss the necessity of these assumptions later.

At the interim stage, it is a dominant strategy for agents to bid their true expected valuation for the good.[4] We look for an equilibrium in which both agents acquire information. It can be shown that, as λ goes to zero, each agent's

[4] As usual in a second-price auction, there are also non-truthful equilibria in weakly dominated strategies. Note, however, that the auction's outcome in such equilibria is independent of bidders' realized valuations. Hence if bidders expected such equilibrium to arise at the interim stage, they would not acquire any information ex ante.

optimal information acquisition strategy leads her to hold one of two posterior beliefs upon entering the auction: one that puts more weight on states in which she has a high valuation $\omega_i = \bar{\omega}_i$, and the other more weight on states where $\omega_i = \underline{\omega}_i$. Let $\overline{m}_i$ be agent i's expected valuation at the former belief, and $\underline{m}_i$ her expected valuation at the latter. So in equilibrium, either agent i learns that state $\bar{\omega}_i$ is more likely, in which case she submits a high bid $\overline{m}_i$, or that $\underline{\omega}_i$ is more likely, in which case she bids $\underline{m}_i$. Naturally, bids need to be consistent with beliefs: If, for instance, agent i always bids $\overline{m}_i$ when $\omega_i = \bar{\omega}_i$ and always bids $\underline{m}_i$ when $\omega_i = \underline{\omega}_i$, then it has to be that agent i is becoming fully informed about her valuation: $\overline{m}_i = \bar{\omega}_i$ and $\underline{m}_i = \underline{\omega}_i$.[5]

Do agents have an incentive to acquire information on the opponent's valuation for the good? We show that the answer is yes, even though the mechanism is strategy-proof at the interim stage, and agents' valuations ω_1, ω_2 are private and independently distributed. To see this, take the perspective of agent 2, and consider what happens if she only learns about her own valuation ω_2. In states where $\omega_1 = \overline{\omega}_1$, her opponent is likely to make a high bid $\overline{m}_1$ ensuring her the object, and learning about her own valuation has no benefit for 2. On the contrary, in states where $\omega_1 = \underline{\omega}_1$, bidder 2's bid impacts the outcome of the auction, and hence information on ω_2 is valuable. Similarly, when $\omega_2 = \underline{\omega}_2$, agent 1 knows that her valuation is higher so does not need acquiring any information on her preferences. When $\omega_2 = \overline{\omega}_2$, she could have a higher or lower valuation than agent 2's most likely bid, and so information on ω_1 is valuable. This shows that the value of information on one's own preferences depends on the realized state for the other. Whether or not agents acquire information on others in equilibrium naturally depends on the cost of information. In particular, agents should not incur a discontinuously high cost upon learning about others, which would offset the associated benefits. Furthermore, agents should

[5]To see why agents only hold two possible posteriors in equilibrium, take the perspective of agent 1. Agent 1 correctly anticipates that agent 2 sends two bids with positive probability in equilibrium. She also knows that she would always want to win the auction if agent 2 were to make a low bid $\underline{m}_2$. (Indeed, a low bid from agent 2 means she believes $\underline{\omega}_2$ to be likely, and so $\underline{m}_2$ is "close" to $\underline{\omega}_2$.) Hence agent 1 needs only figure out whether she wants to win when agent 2 bids high. Her action space then reduces to a binary set (whether or not to outbid $\overline{m}_2$), and an optimal information strategy puts weight on at most two posteriors (one associated with each action.) Similarly, agent 2 needs only figure out whether she wants to outbid $\underline{m}_1$.

not *always* want to become *fully* informed of their own preferences, but should instead equate value and cost of information at the margin.

For the sake of tractability, we characterize what the equilibrium converges to as the scaling parameter λ goes to zero.[6] Agents' equilibrium strategies are summarized in the following table:

	$(\overline{\omega}_1, \overline{\omega}_2)$	$(\overline{\omega}_1, \underline{\omega}_2)$	$(\underline{\omega}_1, \overline{\omega}_2)$	$(\underline{\omega}_1, \underline{\omega}_2)$
$\Pr(\overline{m}_1 \mid \omega_1, \omega_2)$	1	$\frac{1}{2}$	0	$\frac{1}{2}$
$\Pr(\overline{m}_2 \mid \omega_1, \omega_2)$	$\frac{1}{3}$	0	1	0

Agent 1 receives the correct signal, and hence submits the correct bid, whenever the other's valuation is high, as this is when information makes a difference: $\Pr(\overline{m}_1|\overline{\omega}_1, \overline{\omega}_2) \longrightarrow 1$ and $\Pr(\overline{m}_1|\underline{\omega}_1, \overline{\omega}_2) \longrightarrow 0$. Reciprocally, agent 2 submits the correct bid whenever agent 1 submits a low bid with non-zero probability, as this is when information is valuable to 2. An interior probability reflects the fact that the agent is indifferent between the two bids in that state. For instance, in state $(\overline{\omega}_1, \overline{\omega}_2)$, agent 2 is indifferent between bidding high and low as she never wins the auction anyway. Then, optimally, agent 2 does not condition her behavior on this state: $\Pr(\overline{m}_2|\overline{\omega}_1, \overline{\omega}_2) = \Pr(\overline{m}_2) = 0.25 \sum_\omega \Pr(\overline{m}_2|\omega)$ which yields $\Pr(\overline{m}_2|\overline{\omega}_1, \overline{\omega}_2) = 1/3$.[7]

Despite the auction being strategy-proof, each bidder still has an incentive to learn about the opponent's valuation, to assess how much she should learn

[6] This simplifies the analysis of the example as it ensures both agents do in fact acquire information in equilibrium. Following Matějka and McKay (2015), we know that equilibrium strategies follow a logit rule under the entropic cost function:

$$\Pr(\overline{m}_1|\omega_1, \omega_2) = \frac{\Pr(\overline{m}_1)}{\Pr(\overline{m}_1) + \Pr(\underline{m}_1) \exp\left[-\frac{1}{\lambda} \Pr(\overline{m}_2|\omega_1, \omega_2) (\omega_1 - \overline{m}_2) \right]}$$

$$\Pr(\overline{m}_2|\omega_1, \omega_2) = \frac{\Pr(\overline{m}_2)}{\Pr(\overline{m}_2) + \Pr(\underline{m}_2) \exp\left[-\frac{1}{\lambda} \big(1 - \Pr(\overline{m}_1|\omega_1, \omega_2)\big) (\omega_2 - \underline{m}_1) \right]}.$$

These logit rules make the interdependency between both agents' strategies explicit. For instance, the equilibrium probability agent 2 bids $\overline{m}_2$ in state (ω_1, ω_2) is decreasing in the likelihood her opponent will submit a high bid in that state $\Pr(\overline{m}_1|\omega_1, \omega_2)$.

[7] For this to be an equilibrium, it has to be that agent 1 always wants to win the auction when agent 2 bids $\underline{m}_2$, even if she has a low valuation. Using Bayes rule, we know that $\underline{m}_2$ must equal $\underline{\omega}_2 \Pr(\underline{\omega}_2|\underline{m}_2) + \overline{\omega}_2 \Pr(\overline{\omega}_2|\underline{m}_2) = \frac{3}{4}\underline{\omega}_2 + \frac{1}{4}\overline{\omega}_2$, which needs to be below $\underline{\omega}_1$.

about her own valuation. If the agent predicts that the other bidder will submit a high bid, then there are less incentives to learn about one's own preferences, and vice versa. Hence strategy-proofness is not enough to guarantee informational simplicity, which illustrates our main finding: a mechanism is informationally simple if and only if it is *de facto* separable, i.e. agents' reports do not interact with one another to determine the allocation. For instance, in the above example, the seller could offer the good to agent 1 at price $\overline{\omega}_2$, and if 1 refuses, give it to agent 2 at price $\underline{\omega}_2$. This is a dictatorial mechanism, which hence satisfies our separability condition as only agent 1 can influence the outcome, and is informationally simple: agent 1 only wants to learn whether her value is above $\overline{\omega}_2$, and agent 2 does not want to acquire any information at all.

1.2 Setup

Environment. We consider good allocation problems with transferable utilities. A single item needs to be allocated to one of n agents. Let $N = \{1, \dots, n\}$ be the set of agents. There is a finite set of possible states of the world, or fundamentals, that has a product structure $\Omega = \bigtimes_{i \in N} \Omega_i$.[8] Each player's preferences for the good depend on her own fundamental only: $u_i \in \mathcal{U}_i \subseteq \mathbb{R}^{\Omega_i}$, with $\mathcal{U}_i$ being the open set of possible preferences for i. The prior probability distribution $\mu_0 \in \Delta\Omega$ is common knowledge among the players and the designer, and satisfies independence: $\mu_0(\omega) = \prod_{i \in N} \mu_0^i(\omega_i)$ where the superscript μ^i corresponds to the marginal on dimension ω_i. Should players' preferences depend on the entire vector of fundamentals, or should the fundamentals be correlated, we could not make the distinction between player i acquiring information on her preferences or on others'. Our assumptions ensure that statements such as "player i acquires information on player j" have a proper meaning. Moreover, they guarantee that agent i does not have a direct interest in acquiring information on ω_{-i}. Hence in what follows, an agent's payoff will depend on others' fundamentals ω_{-i} *only indirectly* through the mechanism.

Though we state our results for good allocation problems, they easily extend to more general settings in which there is an abstract set of outcomes and

[8] We assume that Ω is finite for simplicity, but it does not appear to drive our results.

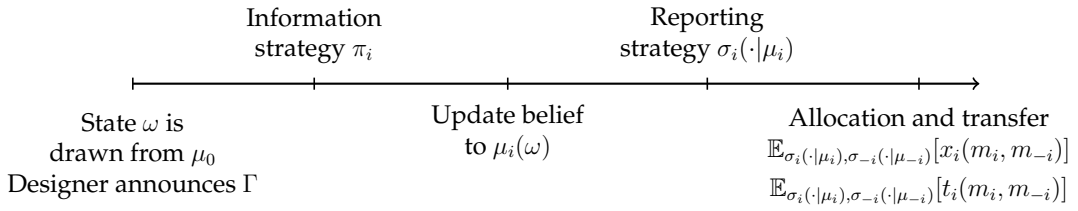


Figure 1.1: *Timing of the game*

agents have arbitrary preferences over these outcomes. The results also extend to environments without transfers (e.g., matching). We focus on good allocation problems and quasi-linear utilities for ease of exposition, and to make our impossibility result stronger as transfers give more leeway to the designer.

Mechanism. A designer ex-ante commits to a mechanism. A mechanism Γ consists of a finite set of messages for each player M_i, as well as allocation functions and transfer functions:

$$x_i : M_1 \times \cdots \times M_n \longrightarrow [0, 1]$$
$$t_i : M_1 \times \cdots \times M_n \longrightarrow \mathbb{R}$$

with $\sum_i x_i(m) \leq 1$ for all $m \in M_1 \times \cdots \times M_n$. The ex-post utility of agent i in state ω under message profile m is quasi-linear in the transfer:

$$x_i(m)u_i(\omega_i) - t_i(m).$$

Strategies. At the ex-ante stage, players can acquire costly information about any state.[9] Information acquisition is represented by choosing a distribution over posterior beliefs $\pi_i \in \Delta\Delta\Omega$ that is consistent with the prior $\mathbb{E}_{\pi_i}[\mu_i] = \mu_0$.[10]

[9]We assume information acquisition is covert for simplicity. None of our results would change if agents could observe the information structure chosen by others before sending their message.

[10]This formulation is equivalent to agents choosing a privately observed signal that can be arbitrarily correlated with the state. Indeed, any signal leads to a specific distribution over posterior beliefs. Reciprocally, any distribution over posteriors π_i satisfying this martingale

At the interim stage, players send a message to the designer conditional on their realized posterior belief μ_i. Let $\sigma_i : \Delta\Omega \longrightarrow \Delta M_i$ denote agent i's reporting strategy. Figure 1.1 describes the timing of the game. Without loss of generality, we directly work with probability distributions over messages conditional on states: $P_i : \Omega \longrightarrow \Delta M_i$. This object is obtained from (π_i, σ_i) using Bayes' rule:[11]

$$P_i(m_i \mid \omega) = \sum_{\mu_i \in \text{supp } \pi_i} \sigma_i(m_i \mid \mu_i) \frac{\mu_i(\omega)}{\mu_0(\omega)} \pi_i(\mu_i).$$

where $\text{supp } \pi_i$ denotes the support of π_i. (Similar notation is used to denote the set of messages in the support of a choice rule P_i.) In words, the information strategy and reporting strategy (π_i, σ_i) lead player i to send message m_i in state ω with probability $P_i(m_i \mid \omega)$. The more an agent's choice rule in state ω differs from that in state ω', the more agent i acquires information to distinguish between states ω and ω'. Conversely, a choice rule P_i that is independent of ω can be implemented without acquiring any information about the state.

A Nash Equilibrium is a strategy profile $(P_i^*)_{i \in N}$ such that, for all $i \in N$,

$$P_i^* \in \underset{P_i \in (\Delta M_i)^\Omega}{\arg\max} \sum_{\omega \in \Omega} \mu_0(\omega) \mathbb{E}_{P_i(\cdot|\omega), P_{-i}^*(\cdot|\omega)} \left[x_i(m_i, m_{-i}) u_i(\omega_i) - t_i(m_i, m_{-i}) \right] - c(P_i)$$

where $c : (\Delta M_i)^\Omega \longrightarrow \mathbb{R}$ is the cost of information acquisition associated with the least informative distribution over posteriors π_i that implements P_i.[12] This

property can be achieved by choosing an appropriate signal (Kamenica and Gentzkow, 2011).

[11] Without loss, an optimal information strategy puts weight on at most $|\Omega|$ posteriors, and so we can assume that the support of π_i is finite. This furthermore implies that considering finite message spaces $(M_i)_i$ is without loss as well.

[12] Any information and reporting strategies (π_i, σ_i) induce a unique choice rule P_i. Conversely, given a choice rule P_i, there exists a least informative distribution over posteriors π_i that implements it, in the sense that all other distributions π_i' that implement P_i are mean preserving spreads of π_i. Intuitively, given any choice rule P_i, there exists a minimal amount of information the agent has to acquire to be able to correlate her reports to the state as specified by P_i. The least informative π_i associated with P_i is derived from Bayes rule in the following way. For all $m_i \in \text{supp } P_i$ let

$$\mu_i^{m_i}(\omega) = \frac{P_i(m_i|\omega)\mu_0(\omega)}{\sum_{\omega'} P_i(m_i|\omega')}$$

be i's posterior when she reports m_i to the designer. Then $\text{supp } \pi_i = \{\mu_i^{m_i} | m_i \in \text{supp } P_i\}$ and $\pi_i(\mu_i^{m_i}) = \sum_\omega P_i(m_i|\omega)$.

implicitly assumes that more information (in Blackwell's order) is costlier, and hence that agents would never buy more information than they need to implement P_i. The above formulation makes clear that information on ω_{-i} can only be valuable to player i if it helps her predict others' equilibrium report $P^*_{-i}(\cdot|\omega)$. Existence of a Nash Equilibrium in pure strategies derives from standard argument, given the assumptions we impose on the cost of information.[13]

Assumptions on the Cost Function. As mentioned above, we assume that more information (in Blackwell's order) is costlier. A choice rule P requires more information to be implemented than P' if P' can be derived by adding noise to P. Formally, P' is a garbling of P if there exists a positive, column-stochastic matrix $[\Lambda_{m_i,m_i'}]_{m_i,m_i'}$ such that $P'(\cdot|\omega) = \Lambda P(\cdot|\omega)$ for all ω.

Assumption 1 (Monotonicity)**.** *If P' is a garbling of P, then $c(P') \leq c(P)$.*

Such monotonicity assumption is standard in the literature. Second, we assume that the cost function is smooth—excluding kinks and jumps,—allowing players to fine-tune the informativeness of signals. A stochastic choice rule is interior if its support is the same across all states: $\operatorname{supp} P_i(\cdot|\omega) = \operatorname{supp} P_i(\cdot|\omega')$ for all ω, ω'. Equivalently, a choice rule is interior if player i's posterior belief always has full support—i.e., player i does not need to rule out some state of the world with certainty in order to implement P_i.

Assumption 2 (Smoothness)**.** *c is twice continuously differentiable and convex over the set of interior stochastic choice rules.*

This ensures that all partial derivatives of c exist and are continuous. In particular, it rules out the possibility that learning about others' preferences is discontinuously costlier than learning about oneself. More generally, there is

[13] As agents' strategy spaces $(\Delta M_i)^\Omega$ are compact, continuity and convexity of the cost c ensure that best-responses are well-behaved—upper hemicontinuous, non-empty, compact and convex valued—by the Theorem of the Maximum. Kakutani's fixed point theorem then guarantees the existence of an equilibrium. Any mixing between two pure strategies P and P' cannot be optimal as it can always be replicated by a pure strategy at lower cost as soon as c is convex. Note that this does not imply that an agent's reporting strategy σ_i never involves some mixing in equilibrium: conditional on some posterior μ_i, an agent might send multiple messages with positive probabilities, but this still translates into a pure choice rule P_i.

no discontinuous change in c, or in the partial derivatives of c, upon learning a bit about others—e.g., the marginal cost associated with sending some message m_i more often in state ω is continuous in P_i, and does not jump when moving from an informationally simple choice rule P_i to one that is not. Absent this assumption, it is easy to find examples in which agents never acquire information on others when utilities are bounded, e.g. take a large fixed cost on acquiring information on ω_{-i} that dominates any benefit from obtaining the good. Since this smoothness assumption is key to our main result, we discuss and relax it in Section 1.3.1.

Third, we assume that the *marginal* cost of information goes to infinity when a player becomes fully informed about any fundamental (though the total cost can be bounded). This Inada condition guarantees that players' optimal choice rules are interior conditional on acquiring information.

Assumption 3 (Inada Condition)**.** *For all $\hat{P}_i$ such that $\hat{P}_i(m_i|\omega) = 0$ and $\hat{P}_i(m_i|\omega') > 0$ for some m_i, ω, ω',*

$$\lim_{P_i \to \hat{P}_i} \frac{\partial c(P_i)}{\partial P_i(m_i|\omega)} = -\infty.$$

In the statement of Assumption 3, $\hat{P}_i$ is a corner choice rule as agent i needs to know with probability one that state ω has not realized in order to implement it. The condition then requires that the marginal cost of this strategy is unbounded, which implies that it is never optimal at equilibrium. Absent this assumption, we can construct environments in which it is always optimal for agents to become *fully* informed about their own preferences, as soon as they have some impact over the outcome. Then, in any strategy-proof mechanism, agents would not have an incentive to acquire information on others.

Finally, we impose that if ω_{-i} is directly and indirectly payoff irrelevant to agent i—think of a dictatorial mechanism where i is the dictator—then player i has no incentive to acquire information on ω_{-i}. This condition is only necessary for the converse of our main result, i.e. to show that a separable mechanism is informationally simple. To define this condition formally, let $V_i(m_i, \omega | P_{-i}, \Gamma) = \mathbb{E}_{P_{-i}(\cdot|\omega)}\Big[x(m_i, m_{-i})u_i(\omega_i) - t_i(m_i, m_{-i})\Big]$ be i's expected payoff from sending

message m_i in state ω given some P_{-i} and mechanism Γ, and denote by P_i^* an optimal choice rule for player i.

Assumption 4 (Independence of Irrelevant States)**.** *For any mechanism Γ and any strategy of others P_{-i}, the following must hold: If $V_i(m_i, (\omega_i, \omega_{-i})|P_{-i}, \Gamma)$ is independent of ω_{-i} for all m_i, then so is $P_i^*(\cdot|\omega_i, \omega_{-i})$.*

In words, if an agent's payoff is independent of some dimensions of the state space, then it is not optimal to learn about these payoff-irrelevant dimensions and the induced optimal choice rule does not depend on them. This relates to other assumptions brought forward in the literature.[14] This assumption in particular rules out the possibility that learning about ω_i is cheaper if one also learns about ω_{-i}.

The most notable example of a cost function satisfying all four conditions is the entropic cost function, which we considered in the motivating example.

Example 1 (Entropic Cost)**.** *Sims (2003) proposes a cost function based on Shannon's entropy which measures a signal's informativeness as the expected reduction in entropy. The entropic cost associated with a stochastic choice rule P_i writes*

$$c(P_i) = -\sum_{m_i \in \text{supp}\, P_i} P_i(m_i) \log P_i(m_i) + \sum_{\omega \in \Omega} \mu_0(\omega) \sum_{m_i \in \text{supp}\, P_i} P_i(m_i|\omega) \log P_i(m_i|\omega),$$

where $P_i(m_i) = \sum_\omega P_i(m_i|\omega)\mu_0(\omega)$ is the unconditional probability of sending message m_i under P_i. Intuitively, the more the agent's choice rule $P_i(\cdot \mid \omega)$ varies across states ω, the higher the cost, as more information about ω is needed to implement it.

We end this section with a comment on static vs. dynamic information acquisition. In our model, agents' choices are static: they simultaneously choose

[14] Independence of irrelevant states is related to but weaker than "invariance under compression" (Caplin et al. (2017)), which is known to hold for the entropic cost function. Invariance under compression requires that splitting a state into two payoff-equivalent states should not change how costly it is to learn about it. In particular, splitting a state ω_i into $|\Omega_{-i}|$ payoff-equivalent states should not change agent i's optimal strategy, which is our above condition. Our condition only requires the cost function to be invariant under compression of some dimensions of the state space (i.e., only of Ω_{-i}), which is reminiscent of a similar condition in Hébert and La'O (2020). Our notion of independence to irrelevant states is conceptually distinct from prior independence of the cost function. In our setting, agents' prior belief μ_0 is fixed, and so whether the cost of information depends on it or not is immaterial.

a choice rule P_i, or equivalently a signal and a reporting strategy. This is, however, not restricting the manner in which agents can acquire information *per se*. Indeed, any dynamic information acquisition process can be reduced to a single, appropriately chosen, signal. For a large class of cost functions, such reduction is without loss as it leads to a weakly lower overall cost of information.[15] This is for instance the case of the entropic cost function, which can be interpreted as the reduced-form expected cost of an optimal binary search tree over the state space.

1.3 The Generic Complexity of Informational Incentives

In this section we address the following question: Which mechanisms provide players with simple informational incentives, i.e. incentives to only acquire information on their own preferences? We show that players have simple informational incentives if and only if the mechanism is *de facto* separable.

Informational simplicity is captured by the following refinement of Nash equilibrium.

DEFINITION 1. *An **Informationally Simple Equilibrium** (ISE) is a Nash equilibrium $(P_i^*)_{i\in N}$ such that, for all i, P_i is independent of ω_{-i}.*[16]

This refinement is of interest for several reasons. First, informational simplicity captures a notion of strategic simplicity which is a priori distinct from strategy-proofness. As it turns out, we will show that informational simplicity is a necessary condition for ex-ante strategy-proofness of the extended game that includes the information acquisition stage. Second, players' interim information structure satisfies the Independent Private value (IPV) assumption if and only if the equilibrium is informationally simple. Hence our analysis shades light on whether we should expect such information structure to arise endogenously.

[15] See Zhong and Bloedel (2020) for a characterization of cost functions satisfying this property.

[16] Equivalently, the signal that agent i acquires is independent of ω_{-i}.

Given others' strategy P^*_{-i}, player i chooses a stochastic choice rule P_i so as to solve the following program:

$$\max_{P_i:\Omega\to\Delta M_i} \quad \sum_{\omega\in\Omega} \mu_0(\omega)\mathbb{E}_{P_i(\cdot|\omega),P^*_{-i}(\cdot|\omega)}\Big[x_i(m_i,m_{-i})u_i(\omega_i) - t_i(m_i,m_{-i})\Big] - c(P_i).$$

Conditional on acquiring some information, the first-order conditions with respect to $P_i(m_i|\omega)\mu_0(\omega)$ yield necessary restrictions on player i's best response: for all m_i in the support of P^*_i and for all $\omega \in \Omega$,

$$\mathbb{E}_{P^*_{-i}(\cdot|\omega)}\Big[x_i(m_i,m_{-i})u_i(\omega_i) - t_i(m_i,m_{-i})\Big] + \frac{\gamma_i(\omega)}{\mu_0(\omega)} = \frac{\partial c(P^*_i)}{\partial P^*_i(m_i|\omega)\mu_0(\omega)}, \qquad (\star)$$

where $\gamma_i(\omega)$ is the Lagrange multiplier associated with the constraint that the choice rule $P_i(\cdot|\omega)$ must sum to one. The left-hand side captures the marginal gain from sending message m_i in state ω (had player i been fully informed of the state), rather than any other message m'_i. The Lagrange multiplier $\gamma_i(\omega)$ is indeed the shadow price of the constraint that the choice rule sums to one, and hence captures the fact that sending m_i more often implies sending $m'_i \neq m_i$ less often. The right-hand side represents the marginal cost associated with sending message m_i more often in state ω, which requires being able to distinguish more often state ω from other states.

To better understand this trade-off, suppose that the cost of any information is very high: then players send only one message that maximizes their average payoff across states. Conversely, if the marginal cost is sufficiently low, then agents send exactly the payoff maximizing message in each state as in a game with perfect information. Hence, for intermediate costs, players achieve a trade-off between the gains from sending the optimal message and the cost associated with discovering what is the optimal message.

We can rearrange and substitute out the Lagrange multiplier to obtain an interpretation of the FOC in terms of value of information. Take the FOC with respect to $P_i(m'_i \mid \omega)\mu_0(\omega)$ and substract from the previous FOC. This gives the marginal gain from reporting m_i relative to m'_i in state ω, net of marginal costs:

$$\mathbb{E}_{P^*_{-i}(\cdot|\omega)}\Big[x_i(m_i,m_{-i})u_i(\omega_i) - t_i(m_i,m_{-i}) - [x_i(m'_i,m_{-i})u_i(\omega_i) - t_i(m'_i,m_{-i})]\Big]$$

$$= \frac{\partial c(P_i^*)}{\partial P_i^*(m_i|\omega)\mu_0(\omega)} - \frac{\partial c(P_i^*)}{\partial P_i^*(m_i'|\omega)\mu_0(\omega)}.$$

In general the value of information for i seems to depend on other agents' reports, as m_{-i} impacts the chosen outcome. In an informationally simple equilibrium, however, player i's strategy must be independent of ω_{-i}. This requires the value of information for player i to be independent of other players' realized state.

We now show that this independence cannot generically be satisfied unless the mechanism is *de facto* separable. A statement holds **generically** if it is false only for a set of utilities $\mathcal{U}^0 \subseteq \bigtimes_i \mathcal{U}_i$ whose closure has Lebesgue measure zero.[17] A mechanism is **separable** if agents' reports cannot interact with one another in the allocation function: for all i, all $m_i, m_i' \in M_i$, and all $m_{-i}, m_{-i}' \in M_{-i}$,

$$x_i(m_i, m_{-i}) - x_i(m_i', m_{-i}) = x_i(m_i, m_{-i}') - x_i(m_i', m_{-i}').$$

Hence others' report m_{-i} can only impact the level in i's outcome, but cannot interact with how i's report affects her outcome. Note that if i cannot influence the outcome altogether, then the condition is trivially satisfied as both sides of the equations are always zero. More generally, it can be that several agents influence the outcome with positive probability, but never *jointly*—e.g., the mechanism might randomly (and independently of reports m) pick a dictator i^* and condition the outcome on her report m_{i^*} only. Finally, say a mechanism is ***de facto* separable** under equilibrium P^* if there exists a separable mechanism $\hat{\Gamma}$ and an equilibrium $\hat{P}^*$ of $\hat{\Gamma}$ that is outcome equivalent.[18]

In most settings of interest, the goal of running a mechanism is precisely to aggregate agents' information, and to choose an allocation based on all pieces of information jointly. Yet we show that this is generically incompatible with informational simplicity.

[17] Mathematical genericity does not necessarily imply genericity "in practice" and very much depends on the chosen universe of preferences. We show that Informational Simplicity is only feasible for preferences that are non-generic, within the open set of allowed preferences $\bigtimes_i \mathcal{U}_i$. But the set of preferences that are relevant in practice could be itself small and non-generic when considering other factors such as the saliency of some strategies.

[18] That is, an equilibrium $\hat{P}^*$ of $\hat{\Gamma}$ leads to the same state-dependent allocation: $\sum_{\hat{m}\in\hat{M}} \prod_i \hat{P}_i^*(\hat{m}_i|\omega)\hat{x}_i(\hat{m}) = \sum_{m\in M} \prod_i P_i^*(m_i|\omega)x_i(m)$ for all i, ω.

THEOREM 1. *Fix any mechanism Γ. Generically, if P^* is an Informationally Simple equilibrium of Γ, then Γ is de facto separable under P^*.*

Conversely, if both the outcome functions and transfer functions of Γ are separable,[19] then all equilibria of Γ are Informationally Simple.

Theorem 1 states that informational simplicity is generically impossible to achieve under most mechanisms of interest. Whenever the mechanism is not de facto separable, it is generically impossible to design transfers that incentivize agents to only learn about themselves. The economic intuition is that players have uncertainty about their "opportunity set," i.e. which outcomes they can bring about. Their opportunity set depends on others' preferences, which makes it valuable to condition how much they learn about their own preferences on the realization of others'. That allows them to acquire more information on their own preferences when the stakes are higher—i.e., when they face a larger opportunity set.[20] Doing so is free at the margin, as the smoothness of the cost function implies it has no discontinuous jump or kink when agents start acquiring information on others. In the proof, we show that this interdependence between the value of information for i and others' preferences is so rich that it cannot be offset by appropriately designed transfers.

Note that under an interim strategy-proof mechanism, agents want to learn about others only because it helps them assess how much they should learn about themselves. Hence it is important for Theorem 1 that agents do not know their own preferences, and that becoming fully informed on their own preferences is never optimal by the Inada condition. It is also essential that agents be able to condition how much they learn about themselves on what they learn about others. This is most intuitive if learning is sequential—e.g., if agents first

[19] That is $t_i(m_i, m_{-i}) - t_i(m'_i, m_{-i}) = t_i(m_i, m'_{-i}) - t_i(m'_i, m'_{-i})$ for all i, m_i, $m'_i \in M_i$, m_{-i}, $m'_{-i} \in M_{-i}$, and similarly for $(x_i)_i$. Note that our baseline definition of separability only imposes restrictions on the outcome functions $(x_i)_i$. This may however not be enough to guarantee Informational Simplicity, as interdependencies in transfers $(t_i)_i$ might incentivize agents to learn about each other. For instance, if transfers generate a coordination game across agents, then agents might coordinate on conditioning their play on one particular agent's state ω_i, which then implies all agents $j \neq i$ learn about another person.

[20] Note that agents do not care about others' preferences ω_{-i} *per se*, but only because it helps them predict others' report to the designer. Hence if agents were allowed to acquire information on what others know—i.e. their posterior beliefs μ_{-i}—as in Denti (2018), then they would do so instead of learning about their underlying preferences ω_{-i}.

buy a signal about ω_{-i} and then, conditional on its realization, buy a signal on ω_i. This is captured by our framework as c can be interpreted as the reduced-form cost of an optimal dynamic process of information acquisition.[21]

Importantly, the necessity part of the theorem is not a statement about the primitives: if P^* is an Informationally Simple equilibrium of Γ then the mechanism *need not be* separable, but under P^* the mechanism acts *as if* it were separable.[22] For instance, when the value of information is very small for all but one player, it is very possible that even if all agents can impact the chosen outcome in the mechanism, only one decides to acquire information in equilibrium. The designer could have then replicated the induced outcome by running a dictatorial mechanism, in which only that agent's private information would have been elicited. That being said, the theorem is sufficient for mechanism design purposes: the point is that informational simplicity is impossible to achieve unless the designer's objective does not require eliciting multiple agents' information and using it jointly to decide on the outcome. Whether or not the mechanism is truly separable or only separable *de facto* is immaterial.

Finally, we emphasize that the assumption of independence of irrelevant state (Assumption 4) only comes into play to prove that a separable mechanism admits an informationally simple equilibrium (sufficiency). Indeed, under a separable mechanism, others' preferences ω_{-i} are (directly and indirectly) payoff-irrelevant to agent i. Hence her choice rule is independent of ω_{-i} only if Assumption 4 holds.

A Knife-Edge Example. We now go through a knife-edge case for which our main result does not hold. Indeed, since the latter is a genericity result, there exists a degenerate set of utility functions under which a mechanism can be both informationally simple and non-separable.

There are three goods $\{A, B, C\}$ to be allocated to two agents $\{i, j\}$.[23] For simplicity, there are no transfers. The mechanism used is a simultaneous ver-

[21] See the discussion at the end of Section 3.

[22] Any separable mechanism is *de facto* separable, but every non-separable mechanisms can also be *de facto* separable.

[23] Knife-edge examples with only one good are more complex and require non-zero transfers, and so we give one with several goods for ease of exposition.

sion of the serial dictatorship: Agents report their preferences, agent i gets her favorite good, then agent j her favorite good among the remaining ones.

Agent i's most preferred good is either A or B: $u_i = (u_{iA}, u_{iB}, u_{iC}) \in \{(4,2,1),(2,4,1)\}$. Agent j always values good A and B equivalently: $u_j \in \{(2,2,0),(0,0,2)\}$. Agent i and j each have two possible messages they can send to the designer: $M_i = \{m_A, m_B\}$ and $M_j = \{m_{AB}, m_C\}$. Intuitively, think of a message m_l as indicating to the designer that the agent wants good l. The allocation function is then:

$x_i(m_i, m_j)$	m_{AB}	m_C
m_A	$(1,0,0)$	$(1,0,0)$
m_B	$(0,1,0)$	$(0,1,0)$

$x_j(m_i, m_j)$	m_{AB}	m_C
m_A	$(0,1,0)$	$(0,0,1)$
m_B	$(1,0,0)$	$(0,0,1)$

As in the motivating example, consider the entropic cost function. We know that optimal choice rules then follow a logit rule (Matějka and McKay (2015)). Since agent i's allocation does not depend on agent j's report, she only learns about her own preferences in equilibrium: the equilibrium probability that i reports m_A in state ω equals

$$P_i^*(m_A|\omega) = \frac{P_i^*(m_A)\exp[\frac{1}{\lambda}u_{iA}(\omega_i)]}{P_i^*(m_A)\exp[\frac{1}{\lambda}u_{iA}(\omega_i)] + P_i^*(m_B)\exp[\frac{1}{\lambda}u_{iB}(\omega_i)]},$$

with $P_i^*(m_l) = \sum_\omega P_i^*(m_l|\omega)$. Note that it is independent of ω_j: agent i does not acquire any information about j's preferences. More surprisingly, agent j also only learns about her own preferences, despite the fact that her allocation is impacted by i's report: the equilibrium probability that j reports m_{AB} in state ω equals

$$P_j^*(m_{AB}|\omega) = \frac{P_j^*(m_{AB})\exp[\frac{1}{\lambda}u_{jA}(\omega_j)]}{P_j^*(m_{AB})\exp[\frac{1}{\lambda}u_{jA}(\omega_j)] + P_j^*(m_C)\exp[\frac{1}{\lambda}u_{jC}(\omega_j)]},$$

which is independent of ω_i.[24] That is because her utility function has some symmetry that makes the value of information on ω_j independent of m_i. Indeed, if agent i picks good A, agent j is left choosing between B and C, and the utility

[24] To derive this logit rule, we use the same formula as for the motivating example: agent j

value of making the correct choice is 2. Similarly, if agent i picks good B, agent j is left choosing between A and C, and the utility value of making the correct choice is, again, 2. The equilibrium is informationally simple, even though two agents acquire information and jointly impact the outcome in equilibrium.

1.3.1 Discussion on Fixed Costs

The proof of Theorem 1 leverages the assumption that the cost of information is smooth. This seems to be a relevant approximation in some settings. For instance, consider a school choice problem in which students must send a rank-order list of schools to a central authority, and can beforehand acquire information on the different schools. They can learn about their own preferences over schools—e.g., by looking at the set of courses offered and whether they look interesting to them—but also about others'—e.g., by asking about the popularity of the school, and admission cutoffs. Arguably, acquiring some information about a school's popularity is not very costly.

A natural concern is that in other settings, the smoothness assumption may be missing relevant factors, and in particular overlooks the possibility that learning about others may be discontinuously harder than learning about oneself. This would mechanically make Informational Simplicity easier to achieve, and we investigate the robustness of Theorem 1 to such discontinuity.

Consider the same smooth cost of information as in our main setup, but suppose that as soon as an agent decides to learn a bit about others, it has to pay an additional fixed cost κ. Fix an arbitrary mechanism Γ and let $\mathcal{U}_{IS}(\kappa) \subseteq \bigtimes_i \mathcal{U}_i$ be the set of utility functions for which (i) there exists an informationally simple equilibrium P^* of Γ, and (ii) Γ is not de facto separable. Let $\rho(\kappa)$ be the Lebesgue measure of $\mathcal{U}_{IS}(\kappa)$.

PROPOSITION 1. *For any mechanism Γ, $\rho(\kappa)$ is increasing and continuous in κ, with $\rho(0) = 0$.*

reports m_{AB} in state ω with probability

$$\frac{P_j^*(m_{AB})\exp[\frac{1}{\lambda}((1-P_i^*(m_A|\omega))u_{jA}(\omega_j)+P_i^*(m_A|\omega)u_{jB}(\omega_j))]}{P_j^*(m_{AB})\exp[\frac{1}{\lambda}((1-P_i^*(m_A|\omega))u_{jA}(\omega_j)+P_i^*(m_A|\omega)u_{jB}(\omega_j))]+P_j^*(m_C)\exp[\frac{1}{\lambda}u_{jC}(\omega_j)]},$$

which simplifies to the above expression since $u_{jB}(\omega_j) = u_{jA}(\omega_j)$ in all states ω_j.

Theorem 1 corresponds to the corner case in which $\kappa = 0$: It is generically impossible to design a mechanism that admits an informationally simple equilibrium and that is non-separable in that equilibrium. As κ increases, the set of preferences for which informational simplicity can be achieved by non-separable mechanisms grows. Interestingly, it grows *continuously*, so our benchmark with $\kappa = 0$ is not a knife-edge case: Adding a small cost to learning about others does make informational simplicity easier to achieve, but only in very few settings. It is only as κ tends to infinity that informational simplicity becomes generically feasible.

1.4 Implications for Mechanism Design

1.4.1 The Limits of Strategy-Proofness

Strategic simplicity is valued in mechanism design for robustness and for leveling the playing field across players. A lot of attention has been given to interim strategy-proof mechanisms, i.e., mechanisms under which agents have a dominant strategy at the interim stage, taking as given their private information. Formally, interim strategy-proofness requires that, for all i and μ_i, there exists $m_i \in M_i$ such that

$$\begin{aligned} &x_i(m_i, m_{-i})\mathbb{E}_{\mu_i}[u_i(\omega_i)] - t_i(m_i, m_{-i}) \\ &\qquad \geq x_i(m_i', m_{-i})\mathbb{E}_{\mu_i}[u_i(\omega_i)] - t_i(m_i', m_{-i}) \quad \forall m_i' \in M_i, m_{-i} \in M_{-i}. \end{aligned}$$

Little is known, however, about the strategic complexity of the acquisition of agents' private information at the ex-ante stage. This is important as many inequalities may arise due to suboptimal information acquisition and strategic mistakes.

Say a mechanism is ex-ante strategy-proof if agents have a dominant strategy in the overall game that includes the information acquisition stage. Formally, for all agent i, there exists a choice rule P_i such that

$$\mathbb{E}_{P_i, P_{-i}}[x_i(m)u_i(\omega_i) - t_i(m)] - c(P_i)$$

$$\geq \mathbb{E}_{P_i', P_{-i}}[x_i(m)u_i(\omega_i) - t_i(m)] - c(P_i') \quad \forall P_i' \in (\Delta M_i)^\Omega, P_{-i} \in (\Delta M_{-i})^\Omega.$$

We show that informational simplicity is a necessary condition for ex-ante dominance solvability.

PROPOSITION 2. *Fix any mechanism Γ. If P^* is an equilibrium in dominant strategy of Γ, then P^* is informationally simple.*

The standard notion of strategy-proofness ensures that agents have a dominant strategy once they have acquired information, but the stronger requirement of Informational Simplicity is needed to guarantee agents also have a dominant strategy when choosing what information to acquire. The intuition behind Proposition 2 is that information about others is valuable only insofar as it helps predict their reports at the interim stage. Hence if an agent learns about another, her equilibrium information strategy has to depend on the strategy of the other player: the mechanism is not ex-ante dominance solvable.

Together, Proposition 2 and Theorem 1 yield that, generically, there exists an equilibrium in dominant strategy of Γ only if Γ is de facto separable. Therefore, in the extended game that includes information acquisition, agents virtually never have a dominant strategy in non-separable mechanisms.

1.4.2 Independent Private Values

A direct corollary of Theorem 1 is that the standard Independent Private value assumption is unlikely to arise endogenously.

Corollary 1. *Fix any mechanism Γ. Generically, the equilibrium posterior beliefs (μ_i, μ_{-i}) are (unconditionally) independent across players only if Γ is de facto separable.*

Therefore the interim information structure is endogenously correlated, which creates interdependent values across players. Put differently, the IPV assumption does not arise endogenously whenever the mechanism is non-separable and the technology of information acquisition satisfies our conditions.

Why has research in mechanism design been limited to the IPV case despite the practical importance of information correlation? A theoretical argument due to Crémer and McLean (1988) suggests that as soon as there is some

correlation in agents' ex-ante private information,[25] the designer can extract all surplus by constructing appropriate side bets. This result highlights that the *independence* of private information across players is necessary for them to earn an information rent as in Myerson (1981). The limits of that result to risk aversion, limited liability, collusion among the agents, etc. have been explored extensively. However what has been explored less is how such results rely on the exogenous nature of private information: If agents anticipate the designer will exploit the correlation structure in their information, why would they acquire such information in the first place? We show that full surplus extraction (in Nash equilibrium) is generically impossible to achieve when taking into account informational incentives. Therefore, our main result together with the impossibility of full surplus extraction suggest that there is room for studying mechanism design with correlated information.

First, we need to properly define what full surplus extraction means in a setting where private information is endogenous. We say full-surplus extraction is feasible if there exists a mechanism that can extract the maximal surplus that can be generated in the economy. Namely, given an environment and a technology of information acquisition, there exists a maximal total surplus that can be generated, which balances total gains from the allocation and total information costs. Full surplus extraction requires that we reach an equilibrium that generates this surplus, and then extract it entirely using transfers.

As in Crémer and McLean (1988), there is one good to be allocated. (Our result easily extends to multiple goods.) Let the **ex-post efficient allocation** at belief profile $\mu = (\mu_i)_{i\in N}$ be the allocation that maximizes total expected welfare:

$$x^*(\mu) = \arg\max_{x\in\Delta N} \sum_{i\in N}\sum_{\omega\in\Omega} \mu_i(\omega)u_i(\omega_i)x_i.$$

The **maximum total surplus** that can be generated in the economy equals:

$$\text{Max. Total Surplus } = \max_{P\in\times_i(\Delta\Omega)^\Omega} \sum_{i\in N}\sum_{\omega\in\Omega} \mu_0(\omega)\mathbb{E}_{P_i,P_{-i}}\left[x_i^*(\mu)u_i(\omega_i)\right] - c(P_i).$$

[25] Formally, whenever the matrix of the conditional probabilities of the signals given the agent types has full rank.

Let $P^\dagger$ be the strategy profile that maximizes total surplus. Note that if $P^\dagger$ is informationally simple—that is, it is socially efficient to have agents acquiring information on themselves only—then using side bets to extract all surplus is trivially precluded. However, we know from Theorem 1 that this is generically not the case whenever the ex-post efficient allocation is non-separable on the support of $P^\dagger$.[26] Hence, whenever efficiency requires that multiple agents learn about their valuations for the good, $P^\dagger$ is generically not informationally simple: it is more efficient for an agent to condition her learning about herself on others' valuations so as to save on information costs whenever possible. In most settings of interest, the information structure that maximizes total surplus then exhibits interdependent beliefs across agents, and allows in principle for the possibility of side bets à la Crémer McLean. We however show that extracting all surplus is generically infeasible, as the anticipation of such side bets distorts agents' incentive to acquire information ex ante.

A (direct revelation) mechanism[27] **extracts the full surplus** if it induces an equilibrium P^* such that:

$$\sum_{i\in N}\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i^*,P_{-i}^*}[t_i(\mu_i,\mu_{-i})] = \text{ Max. Total Surplus},$$

while satisfying incentive and individual rationality constraints. Incentive constraints are of two sorts here: agents should be incentivized to reveal their private information to the designer at the interim stage, and should find it optimal to acquire the socially efficient level of information at the ex-ante stage. The former is the standard IC constraint in mechanism design, and from now on, suppose it holds. The latter, which is the one limiting the possibility of full

[26] Indeed, if $P^\dagger$ is informationally simple, then we can design a mechanism which is informationally simple and not de facto separable, by setting $M_i = \operatorname{supp} P_i^\dagger$, $x(\mu_i,\mu_{-i}) = x^*(\mu)$ and $t_i(\mu_i,\mu_{-i}) = -\sum_{j\neq i}\sum_\omega \mu_j(\omega)x_j^*(\mu)u_j(\omega_j)$. Such a mechanism can however exist only for a non-generic set of preferences.

[27] The standard Revelation Principle applies in our setting: If $(P_i^*)_{i\in N}$, or equivalently $(\pi_i^*,\sigma_i^*)_{i\in N}$, is an equilibrium of Γ then there is an outcome-equivalent direct revelation mechanism $\widehat{\Gamma}$ in which the principal elicits agents' beliefs $\widehat{M}_i = \operatorname{supp}\pi_i^*$ and commits to implementing their equilibrium strategy $(\hat{x}(\mu_i,\mu_{-i})),\hat{t}(\mu_i,\mu_{-i})) = (x(\sigma^*(\mu_i,\mu_{-i})),t(\sigma^*(\mu_i,\mu_{-i})))$.

surplus extraction in this setting, writes:

$$\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i^\dagger,P_{-i}^\dagger}\left[x_i^*(\mu)u_i(\omega_i)-t_i(\mu_i,\mu_{-i})\right]-c(P_i^\dagger)$$
$$\geq\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i,P_{-i}^\dagger}\left[x_i^*(\mu)u_i(\omega_i)-t_i(\mu_i,\mu_{-i})\right]-c(P_i)\quad\text{for all } P_i,i.$$

Finally, the mechanism should satisfy the following ex-ante and interim individual rationality constraints:

$$\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i^\dagger(\cdot|\omega),P_{-i}^\dagger(\cdot|\omega)}\left[x_i^*(\mu)u_i(\omega_i)-t_i(\mu_i,\mu_{-i})\right]-c(P_i^\dagger)\geq 0\quad\text{for all } i.$$
$$\sum_{\omega\in\Omega}\mu_i(\omega)\mathbb{E}_{P_{-i}^\dagger(\cdot|\omega)}\left[x_i^*(\mu)u_i(\omega_i)-t_i(\mu_i,\mu_{-i})\right]\geq 0\quad\text{for all } i,\mu_i\in\operatorname{supp}P_i^\dagger.$$

Hence to extract the full surplus, the mechanism must (i) induce agents to acquire the socially efficient level of information, (ii) pick the ex-post efficient allocation given reported posterior beliefs, and (iii) have transfers that extract all surplus net of information acquisition costs. The last two requirements are familiar from Crémer and McLean (1988), whereas the first one is new but necessary to make sense of ex-post efficiency.

Observe that in some extreme cases full surplus extraction is possible. For instance, consider a setting in which the efficient allocation is the same in every state ω. This means that $x^*(\mu)=x^*$ is independent of agents' posterior beliefs and that the efficient information strategy is to acquire no information at all. The mechanism that always selects outcome x^* irrespective of agents' reports, and has transfers $t_i=\sum_\omega\mu_0(\omega)x_i^*u_i(\omega_i)$ is individually rational, incentive compatible, and extracts full surplus.

However informational incentives generically limit the possibility of full surplus extraction whenever it is efficient for agents to acquire some information. To prove this, we show that the three requirements exposed above translate into necessary conditions that are generically mutually incompatible. We first focus on requirements (i) and (ii) of full surplus extraction, namely that the mechanism induces socially efficient information acquisition and implements the ex-post efficient allocation. Conditional on $P_{-i}^\dagger$, agent i's optimal strategy

solves:

$$\max_{P_i \in (\Delta\Omega)^\Omega} \sum_{\omega \in \Omega} \mu_0(\omega) \mathbb{E}_{P_i, P_{-i}^\dagger} \left[x_i^*(\mu) u_i(\omega_i) - t_i(\mu_i, \mu_{-i}) \right] - c(P_i).$$

The standard approach to incentivize efficient information acquisition is to use Clarke pivot rule:

$$t_i(\mu) = - \sum_{j \neq i} \sum_{\omega \in \Omega} \mu_j(\omega) x_j^*(\mu) u_j(\omega_j).$$

We show that such transfers are actually the *only* one inducing agents to acquire the socially efficient level of information. This result is reminiscent of Hatfield et al. (2018) who extend the Green–Laffont–Holmström theorem by showing that VCG mechanisms with ex-ante costly investments are the unique efficient and strategy-proof mechanisms. To obtain a Crémer-McLean mechanism and enforce the third requirement of full surplus extraction, we add side-bets $b_i : \bigtimes_{j \neq i} \Delta\Omega_j \longrightarrow \mathbb{R}$. Therefore the transfers write:

$$t_i^{CM}(\mu) = - \sum_{j \neq i} \sum_{\omega \in \Omega} \mu_j(\omega) x_j^*(\mu) u_j(\omega_j) + b_i(\mu_{-i}).$$

Such side bets, however, generically distort *informational* incentives, that is incentives to acquire the efficient level of information. This reduces the total surplus generated by the mechanism, preventing full surplus extraction.

THEOREM 2. *Suppose that it is socially efficient for at least one agent to acquire some information. Then full surplus extraction is generically infeasible.*

The feasibility of full surplus extraction with information acquisition received mixed answers in the literature. For instance, Bikhchandani (2010) shows that when the set of signals agents can acquire on others' type is small enough, then full surplus extraction is feasible. Instead, when the set of signals is large enough, then full surplus extraction becomes impossible. Our result confirm that when information acquisition is sufficiently flexible (in our case, fully flexible) then full surplus extraction seems impossible.

In this section, we took an *ex-ante* perspective to full surplus extraction, re-

quiring that the mechanism extracts the maximal surplus that can be generated in the economy. Another approach would be to ask whether *ex-post* full surplus extraction is possible: Does there exist a mechanism such that, in equilibrium, the ex-post efficient allocation is implemented and all the associated surplus is extracted from agents? Here the answer is always yes: the constant mechanism that always picks the efficient allocation at the prior $x^*(\mu_0)$ and has transfers equal to $t_i = \sum_\omega \mu_0(\omega) x_i^*(\mu_0) u_i(\omega_i)$ induces no information acquisition, and does extract all ex-post surplus in equilibrium. Even if we restrict attention to equilibria in which agents acquire some information, it seems to be always possible to find a mechanism extracting all ex-post surplus in equilibrium. This, however, is no guarantee on the magnitude of the surplus that is extracted by the seller: it can very well be that the generated surplus in equilibrium is very small.

1.5 Conclusion

In this paper, we investigate players' *informational incentives* in mechanism design, namely how the choice of the mechanism impacts what information players acquire in equilibrium. A mechanism is informationally simple if players have no incentives to acquire information on others' preferences. Our main result is that, for any smooth technology of information acquisition satisfying an Inada condition, a mechanism is Informationally Simple if and only if it is *de facto* separable. Separability means that agents' report cannot interact with one another in the allocation function, which rules out most economically meaningful mechanisms. This result holds generically, that is for an open set of preferences that has full measure. The intuition is that the outcomes a player can bring about in a mechanism depend on others' report, which makes it optimal to acquire information on them before investing in information acquisition on her own preferences.

This result has two implications for mechanism design. First, we show that a mechanism is ex-ante dominance solvable only if it has an informationally simple equilibrium, hence only if the mechanism is *de facto* separable. This points to a limitation of strategy-proofness as a concept of strategic simplic-

ity. Indeed, even interim strategy-proof mechanisms incentivize players to acquire information about others' and to best respond to beliefs about opponents' play at the ex-ante stage. Second, our result suggests that the independent private value assumption is unlikely to arise endogenously. This, however, does not mean full surplus extraction is possible using side bets as in Crémer and McLean (1988), as these would distort players' incentives when acquiring information.

There are several avenues for future research. One source of information acquisition that we ignored is communication among players. On one hand, our result suggests that some players would benefit from information aggregation in a communication stage after the information acquisition stage, as players endogenously hold information relevant to others. On the other hand, adding such a communication stage would modify informational incentives and free-riding may arise in the information acquisition stage. This raises an interesting question: Under what conditions does communication facilitate implementation and would arise endogenously from a coalition of players?

These considerations suggest that informational incentives may have important and concrete implications for the design of institutions—which remain largely unexplored to this day.

Proofs

Proof of Theorem 1. We start by the proof of necessity, which is more involved than that of sufficiency. By contradiction, suppose that there exists an IS equilibrium P^* of Γ but Γ is not de facto separable under P^*.

First we show that, because Γ is not de facto separable, at least two players must acquire information in equilibrium. By definition, if Γ is not de facto separable, then there exist no mechanism $\widehat{\Gamma}$ that is separable and induces the same state-dependent outcome as Γ. In particular, the direct revelation mechanism $\widehat{\Gamma}$ associated with equilibrium P^* of Γ is non-separable. Let $M_i^* \equiv \{m_i \mid \sum_\omega P_i^*(m_i|\omega) > 0\}$ be the set of messages i sends with positive probability in equilibrium and $\mu_i^{m_i} \equiv (P_i(m_i|\omega_i)\mu_0(\omega_i))/(\sum_{\omega_i'} P_i^*(m_i|\omega_i'))$ her belief when she sends m_i. The direct revelation mechanism asks agents to report their equi-

librium beliefs $\widehat{M}_i \equiv \{\mu_i^{m_i} | m_i \in M_i^*\}$ and implements the same outcome as Γ: $\hat{x}(\mu_i^{m_i}, \mu_{-i}^{m_{-i}}) = x(m_i, m_{-i})$. For this mechanism not to be separable, there must exist an agent i such that $|\widehat{M}_i| \geq 2$ and $|\widehat{M}_{-i}| \geq 2$, and

$$\hat{x}_i(\mu_i^{m_i}, \mu_{-i}^{m_{-i}}) - \hat{x}_i(\mu_i^{m_i'}, \mu_{-i}^{m_{-i}}) \neq \hat{x}_i(\mu_i^{m_i}, \mu_{-i}^{m_{-i}'}) - \hat{x}_i(\mu_i^{m_i'}, \mu_{-i}^{m_{-i}'})$$

for some $\mu_i^{m_i}, \mu_i^{m_i'} \in \widehat{M}_i$, $\mu_{-i}^{m_{-i}}, \mu_{-i}^{m_{-i}'} \in \widehat{M}_{-i}$. This has several implications. First, it must be that $\mu_i^{m_i} \neq \mu_i^{m_i'}$, i.e. that agent i's belief when reporting m_i is different from her belief when reporting m_i' in equilibrium P_i^*. This means $P_i^*(m_i|\cdot) \neq P_i^*(m_i'|\cdot)$ and ensures that i does acquire some information in equilibrium. Similarly, it must be that $\mu_{-i}^{m_{-i}} \neq \mu_{-i}^{m_{-i}'}$ and hence that $P_{-i}^*(m_{-i}|\cdot) \neq P_{-i}^*(m_{-i}'|\cdot)$. This ensures that other agents also acquire information, and that the way they do so impacts how much agent i can influence the outcome. From now on, we focus on the incentives of this particular agent i.

Second, we show that, for almost all preferences of i in $\mathcal{U}_i$, i's optimal strategy is not informationally simple. That is, generically, there does not exist transfers $t_i \in \mathbb{R}^{M^*}$ such that the strategy P_i^* that solves i's system of FOCs ($\star$) is informationally simple. Since what matters for agent i is how her preferences compare from one state to another, we fix agent i's preferences in some arbitrarily chosen state $u_i(\omega_i^0)$ and show that for almost all $(u_i(\omega_i))_{\omega_i \neq \omega_i^0} \in \mathcal{U}_i^{-u_i(\omega_i^0)} \equiv \{(u_i(\omega_i))_{\omega_i \neq \omega_i^0} | (u_i(\omega_i^0), (u_i(\omega_i))_{\omega_i \neq \omega_i^0}) \in \mathcal{U}_i\}$, i's optimal strategy is not IS. To do so, consider the FOCs ($\star$) corresponding to agent i and messages in M_i^*. Since we know that these messages are sent with positive probability in equilibrium, we can ignore the non-negativity constraints on equilibrium probabilities. By the Inada condition we furthermore know that the equilibrium stochastic choice rule P_i^* must be interior, and hence that these FOCs must hold with equality. Note that the endogenous variables in the FOCs are not only the agent's choice rule P_i but also the Lagrange multipliers γ_i. To avoid carrying the multipliers around in the analysis, we substitute them out by choosing an arbitrarily message $m_i^0 \in M_i^*$, and substracting the FOC for message m_i^0 to the FOCs for messages $m_i \in M_i^* \setminus \{m_i^0\}$. Let $\mathcal{P} \equiv (\Delta(M_i^*))^{\Omega_i} \times \mathbb{R}^{M^*}$ and define $\Phi : \mathcal{P} \times \mathcal{U}_i^{-u_i(\omega_i^0)} \longrightarrow \mathbb{R}^{(|M_i^*|-1)\times|\Omega|}$ as the function that maps stochastic choice rules with support M_i^*, and transfers $(P_i, t_i) \in \mathcal{P}$ together with preferences

$u_i \in \mathcal{U}_i^{-\omega_i^0}$ to the following vector:

$$\mathbb{E}_{P^*_{-i}(\cdot|\omega)}\Big[(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i}))u_i(\omega_i) - (t_i(m_i, m_{-i}) - t_i(m_i^0, m_{-i}))\Big] \\ - \frac{\partial c(P_i)}{\partial P_i(m_i|\omega_i)\mu_0(\omega)} + \frac{\partial c(P_i)}{\partial P_i(m_i^0|\omega_i)\mu_0(\omega)}$$

for all $m_i \in M_i^* \setminus \{m_i^0\}, \omega \in \Omega$. Substituting out the Lagrange multipliers from $(\star)$ also makes it clear that we can normalize i's transfers associated with one particular message, for instance $t_i(m_i^0, \cdot)$, as only the relative payoff between sending one message instead of another matters for i's optimal strategy. The vector of transfers is then effectively an element of $\mathbb{R}^{(|M_i^*|-1)\times|M_{-i}^*|}$. More importantly, i's stochastic choice rule is informationally simple by assumption, and thus belongs to $\mathbb{R}^{(|M_i^*|-1)\times|\Omega_i|}$. Therefore we have

$$\dim \mathcal{P} = (|M_i^*| - 1) \times (|\Omega_i| + |M_{-i}^*|).$$

The FOCs for agent i can be written as $\Phi(P_i^*, t_i; u_i) = \mathbf{0}$. Hence, the set of IS stochastic choice rules (together with transfers) which solve the agent's FOCs is $\Phi^{-1}(\mathbf{0}; u_i)$. We show that this set is a manifold of negative dimension, and hence is empty, for almost all $u_i \in \mathcal{U}_i^{-u_i(\omega_i^0)}$. Since this is true irrespective of the normalization we choose for $u_i(\omega_i^0)$ and because $\mathcal{U}_i = \bigcup_{\omega_i^0} u_i(\omega_i^0) \times \mathcal{U}_i^{-u_i(\omega_i^0)}$, this implies that there exists no IS solution to i's system of FOCs for almost all preferences in i's overall set of possible preferences $\mathcal{U}_i$. This is done by successively applying the Transversality theorem (to show that the non-linear equations in this system are locally linearly independent at $\mathbf{0}$ for almost all u_i), and the Regular Value theorem (to show that the solution set is a manifold of negative dimension).[28]

In order to apply the Transversality theorem we need to show that $\mathbf{0}$ is a regular value of Φ, i.e. that the Jacobian of Φ at $\mathbf{0}$ has full rank: $\Phi(P_i, t_i; u_i) = \mathbf{0} \implies \operatorname{rank} D\Phi(P_i, t_i; u_i) = \min\{(|M_i^*| - 1) \times |\Omega|, \dim \mathcal{P} + \dim \mathcal{U}_i^{-u_i(\omega_i^0)}\}$ where D is the Jacobian. Intuitively, this is equivalent to showing that the number of

[28] Mas-Colell (1989) Chapter 1 (section H) and especially Chapter 8 provide an introduction to differential topology. A formal statement of the results we use here can be found on page 320.

locally linearly independent equations of the system evaluated at $\mathbf{0}$ is maximal. Note that $D\Phi$ has $(|M_i^*|-1)\times|\Omega|$ rows—one for each FOC, so one for each $m_i \in M_i^* \setminus \{m_i^0\}$ and $\omega \in \Omega$—and $\dim \mathcal{P} + \dim \mathcal{U}_i^{-u_i(\omega_i^0)}$ columns—each corresponding to the derivative of Φ with respect to one element of $(P_i, t_i; u_i)$. We show that the columns of $D\Phi$ are linearly independent.

The Jacobian of Φ has some simplifying structure, as many of its entries are zero. First, the columns associated with the derivatives w.r.t. P_i correspond to the Hessian of the cost of information, as P_i only enter the FOCs through the marginal cost:[29]

$$D_{P_i}\Phi = \left[\frac{\partial^2 c(P_i)}{\partial P_i(m_i^0|\omega)\mu_0(\omega)\partial P_i(m_i'|\omega')\mu_0(\omega')} - \frac{\partial^2 c(P_i)}{\partial P_i(m_i|\omega)\mu_0(\omega)\partial P_i(m_i'|\omega')\mu_0(\omega')}\right]_{((m_i,\omega),(m_i',\omega'))}.$$

Second, since $(t_i(m_i, m_{-i}))_{m_{-i}}$ only enter the FOCs of agent i associated with sending message m_i, the columns associated with the derivatives w.r.t. t_i form a block diagonal matrix with each block corresponding to one message m_i for agent i:

$$D_{t_i}\Phi = \begin{bmatrix} B_{t_i}(m_i) & \mathbf{0} & \dots & \\ \mathbf{0} & \ddots & & \\ \vdots & & B_{t_i}(m_i') & \mathbf{0} \\ & & \mathbf{0} & \ddots \end{bmatrix}$$

In a block $B_{t_i}(m_i)$, each row corresponds to a possible state (ω_i, ω_{-i}). The

[29] D_{P_i} denotes the restriction of the Jacobian corresponding to the derivative w.r.t. P_i.

columns correspond to derivatives w.r.t. $(t_i(m_i, m_{-i}))_{m_{-i}}$:

$$B_{t_i}(m_i) = \begin{bmatrix} -P_{-i}(m_{-i}|\omega_{-i}) & -P_{-i}(m'_{-i}|\omega_{-i}) & \dots \\ -P_{-i}(m_{-i}|\omega'_{-i}) & -P_{-i}(m'_{-i}|\omega'_{-i}) & \\ & & \ddots \\ -P_{-i}(m_{-i}|\omega_{-i}) & -P_{-i}(m'_{-i}|\omega_{-i}) & \dots \\ -P_{-i}(m_{-i}|\omega'_{-i}) & -P_{-i}(m'_{-i}|\omega'_{-i}) & \\ & & \ddots \end{bmatrix}$$

Similarly, since each $u_i(\omega_i)$ only enters the FOCs of agent i in state ω_i, the columns associated with its derivative have non-zero entries only for rows that correspond to FOCs in state ω_i. For these rows, the derivative equal

$$\sum_{m_{-i} \in M^*_{-i}} P_{-i}(m_{-i} \mid \omega_{-i})(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i})).$$

We first argue that the columns of D_{P_i,t_i} are linearly independent. Note that the columns corresponding to derivatives w.r.t. t_i give the probability that others send m_{-i} conditional on ω_{-i}, for each m_{-i}. Importantly these probabilities are independent of ω_i by assumption, and thus are constant across rows that differ only by ω_i. On the contrary, $\partial^2 c(P_i^*)/\partial P_i^*(m_i|\omega_i)\mu_0(\omega)^2$ varies with ω_i since i acquires information in equilibrium, and it is thus impossible to express the first sets of columns (corresponding to derivatives w.r.t. P_i) in terms of the second (corresponding to derivatives w.r.t. t_i). Furthermore, the columns corresponding to derivatives w.r.t. t_i are also linearly independent as we know other agents $-i$ acquire some information in equilibrium. Hence it cannot be that the likelihood they send some message m_{-i} in each state is the same as for some other message m'_{-i}, as that would mean they hold the same belief when sending m_{-i} and m'_{-i}.

We now show that the columns of D_{u_i} are linearly independent from those of D_{P_i,t_i}. Using a similar argument as above, derivatives w.r.t. u_i must be linearly independent from those w.r.t. P_i as the former depend on ω_{-i} whereas the latter do not. Indeed, and as discussed above, the fact that Γ is not de facto

non-separable implies $\sum_{m_{-i}\in M^*_{-i}} P_{-i}(m_{-i} \mid \omega_{-i})(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i}))$ must vary with ω_{-i}. The main thing to prove is that the columns of D_{u_i} are linearly independent from the columns corresponding to derivatives w.r.t. t_i. Recall that only $(u_i(\omega_i))_{\omega_i \neq \omega_i^0}$ are parameters here, as $u_i(\omega_i^0)$ is normalized to some fixed and arbitrary value. Hence all rows corresponding to FOCs in state ω_i^0 must have zero entries in $D_{u_i}\Phi$. All other entries equal $\sum_{m_{-i}\in M^*_{-i}} P_{-i}(m_{-i} \mid \omega_{-i})(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i}))$, and could be replicated using the $D_{t_i}\Phi$ columns by weighting by $x_i(m_i, m_{-i})$ the column corresponding to the derivative w.r.t. $t_i(m_i, m_{-i})$. However, this is not possible as it would need to generate a zero entry for state ω_i^0 which is possible only if $\sum_{m_{-i}\in M^*_{-i}} P_{-i}(m_{-i} \mid \omega_{-i})(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i})) = 0$ for all m_i, which cannot be true in a non-separable mechanism.

Thus, if Γ is not de facto separable, $D\Phi(P_i, t_i; u_i)$ has full rank and $\mathbf{0}$ is a regular value of Φ. The Parametric Transversality theorem states that, except for a nullset $\overline{\mathcal{U}}_i^{-u_i(\omega_i^0)} \subset \mathcal{U}_i^{-u_i(\omega_i^0)}$ of preferences, $\mathbf{0}$ is a regular value of $\Phi(\cdot; u_i)$. Then by the Regular Value theorem, $\Phi^{-1}(\mathbf{0}; u_i)$ is a smooth manifold of dimension

$$\dim \Phi^{-1}(\mathbf{0}; u_i) = (|M_i^*| - 1) \times (|\Omega_i| + |M_{-i}^*|) - (|M_i^*| - 1) \times |\Omega_i| \times |\Omega_{-i}| < 0.$$

The inequality comes from Blackwell's principle of irrelevant information, which implies $|M_{-i}^*| \leq |\Omega_{-i}|$ in any IS equilibrium, as information is valuable only insofar as it changes the optimal action. Therefore we conclude that, for a full measure set of preferences $\mathcal{U}_i^{-u_i(\omega_i^0)} \setminus \overline{\mathcal{U}}_i^{-u_i(\omega_i^0)}$, the set of IS stochastic choice rules (together with transfers) solving the FOCs is empty. Let $\mathcal{U}_i^0 \equiv \cup_{\omega_i^0} u_i(\omega_i^0) \times \overline{\mathcal{U}}_i^{-u_i(\omega_i^0)}$ be the overall set of preferences for i for which there is an IS solution to i's system of FOCs. Since $\overline{\mathcal{U}}_i^{-u_i(\omega_i^0)}$ has Lebesgue measure zero for each possible normalization of $u_i(\omega_i^0)$, the overall set of preferences $\mathcal{U}_i^0 \subset \mathcal{U}_i$ for which i has an informationally simple optimal strategy is null as well.

We have left to show that $\mathcal{U}_i^0$ is closed, or equivalently that $\mathcal{U}_i \setminus \mathcal{U}_i^0$ is open. Take any $u_i \in \mathcal{U}_i \setminus \mathcal{U}_i^0$. By definition, for these preferences, there does not exist transfers that make i's optimal strategy Informationally Simple. That is, there does not exist (P_i, t_i) such that $\Phi(P_i, t_i; u_i) = 0$. Let $\|\cdot\|$ denote the Euclidean distance, and note that the minimum of $\|\Phi(\cdot; u_i)\|$ is reached for some $(P_i, t_i) \in \mathcal{P}$. Indeed, any large enough t_i or boundary choice rule P_i send $\|\Phi(\cdot; u_i)\|$ to infin-

ity, and so we can restrict attention to a compact subset of $\mathcal{P}$ to find a minimizer of $\|\Phi(\cdot; u_i)\|$. Since $\|\Phi(\cdot; u_i)\|$ is continuous on such compact subset, it must reach a minimum. Let $\delta \equiv \min_{(P_i,t_i)} \|\Phi(P_i, t_i; u_i)\|$, with $\delta > 0$ by assumption. Take any $\varepsilon \in (0, \delta(|M_i^*|)^{-1/2})$, and consider any $u_i' \in \mathcal{U}_i$ such that $\|u_i - u_i'\| < \varepsilon$. Then, for any (P_i, t_i),

$$\begin{aligned}
&\|\Phi(P_i, t_i; u_i) - \Phi(P_i, t_i; u_i')\| \\
&\qquad = \left(\sum_{\omega, m_i} (\mathbb{E}_{P_{-i}(\cdot|\omega)}[(x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i}))(u_i(\omega_i) - u_i'(\omega_i)])^2 \right)^{\frac{1}{2}} \\
&\qquad = \left(\sum_{\omega} \left(\sum_{m_i} \mathbb{E}_{P_{-i}(\cdot|\omega)}[x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i})]^2 \right) (u_i(\omega_i) - u_i'(\omega_i)^2 \right)^{\frac{1}{2}} \\
&\qquad \leq \sqrt{|M_i^*|} \left(\sum_{\omega} (u_i(\omega_i) - u_i'(\omega_i)^2 \right)^{\frac{1}{2}} < \sqrt{|M_i^*|}\varepsilon,
\end{aligned}$$

where the inequality follows from $\mathbb{E}_{P_{-i}(\cdot|\omega)}[x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i})] \leq 1$, and thus $\sum_{m_i} \mathbb{E}_{P_{-i}(\cdot|\omega)}[x_i(m_i, m_{-i}) - x_i(m_i^0, m_{-i})]^2 \leq |M_i^*|$. This implies $\|\Phi(P_i, t_i; u_i')\| > \delta - \varepsilon > 0$, and $u_i' \in \mathcal{U}_i \setminus \mathcal{U}_i^0$. Hence $\mathcal{U}_i \setminus \mathcal{U}_i^0$ is open, and the system of FOCs for i has an IS solution only for a set of preferences $\mathcal{U}_i^0$ whose closure has Lebesgue measure zero.

We now prove that if both the outcome and transfer functions of Γ are separable, then all equilibria of Γ are Informationally Simple. Outcome and transfer functions being separable means that the way an agent i impacts her outcome/transfer only depends on her messages. Formally, there exist mappings $X_i : M_i \times M_i \longrightarrow [0, 1]$ and $T_i : M_i \times M_i \longrightarrow \mathbb{R}$ for all i such that

$$\begin{aligned}
x_i(m_i, m_{-i}) - x_i(m_i', m_{-i}) &= X_i(m_i, m_i') \\
t_i(m_i, m_{-i}) - t_i(m_i', m_{-i}) &= T_i(m_i, m_i')
\end{aligned}$$

for all m_{-i}. Consider some agent i, who takes as given others' strategy P_{-i}^*. Her

objective is

$$\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i(\cdot|\omega),P^*_{-i}(\cdot|\omega)}\Big[x_i(m_i,m_{-i})u_i(\omega_i)-t_i(m_i,m_{-i})\Big]-c(P_i).$$

Since in each state ω her choice rule must sum to one $\sum_{m_i}P_i(m_i|\omega)=1$, we can normalize agent i's utility by her expected utility from sending some arbitrarily chosen message $m_i^0\in M_i$:

$$\begin{aligned}\mathbb{E}_{P_i,P^*_{-i}}\Big[(x_i(m_i,m_{-i})-x_i(m_i^0,m_{-i}))u_i(\omega_i)-(t_i(m_i,m_{-i})-t_i(m_i^0,m_{-i}))\Big]\\ -\mathbb{E}_{P^*_{-i}}\Big[x_i(m_i^0,m_{-i})u_i(\omega_i)-t_i(m_i^0,m_{-i})\Big]-c(P_i),\end{aligned}$$

So what matters for agent i is the relative payoff she gets under the different messages she can send. Since the mechanism is separable, her objective can be equivalently expressed as

$$\sum_{\omega\in\Omega}\mu_0(\omega)\mathbb{E}_{P_i(\cdot|\omega)}\left[X_i(m_i,m_i^0)u_i(\omega_i)-T_i(m_i,m_i^0)\right]-c(P_i).$$

This formulation makes it clear that the value agent i gets from sending message m_i in state $\omega=(\omega_i,\omega_{-i})$ only depends on ω_i and not on ω_{-i}. By Assumption 4, agent i's optimal choice rule must be independent of payoff-irrelevant states: $P_i^*(\cdot|\omega_i,\omega_{-i})=P_i^*(\cdot|\omega_i,\omega'_{-i})$ for all ω_i, ω_{-i}, ω'_{-i}. This holds for all agents, and thus all equilibria of Γ must be informationally simple. □

Proof of Proposition 1. Let $\mathcal{U}_{IS}(\kappa)\subseteq\bigtimes_i\mathcal{U}_i$ be the set of utility functions for which (i) there exists an IS equilibrium P^* of Γ, and (ii) Γ is not de facto separable.

The case with $\kappa=0$ is the baseline case considered in this paper, for which Theorem 1 applies: $\mathcal{U}_{IS}(0)$ has Lebesgue measure zero, hence $\rho(0)=0$.

To show that ρ is increasing, we prove that for any κ, κ' with $\kappa'\geq\kappa$, $\mathcal{U}_{IS}(\kappa)\subseteq\mathcal{U}_{IS}(\kappa')$. Take any $u\in\mathcal{U}_{IS}(\kappa)$. By definition, we know that there exists a non-separable IS equilibrium P^*. We need to show that P^* remains an equilibrium if we increase the fixed cost from κ to κ'. Now that we have introduced a discontinuity in the objective function of agents, the FOCs ($\star$) are not sufficient to characterize an equilibrium. There are two possible types of equilibrium strategies for an agent: either she learns about others or not. If she does, then

her strategy must satisfy ($\star$). If she does not, then her IS strategy must solve:

$$\mathbb{E}_{\mu_0^{-i}}\left(\mathbb{E}_{P^*_{-i}(\cdot|\omega_{-i})}\Big[x_i(m_i,m_{-i})u_i(\omega_i)-t_i(m_i,m_{-i})\Big]+\frac{\gamma_i(\omega)}{\mu_0(\omega)}-\frac{\partial c(P_i^*)}{\partial P_i^*(m_i|\omega)\mu_0(\omega)}\right)=0$$

for all ω_i and all $m_i \in \operatorname{supp} P_i^*$. These two sets of FOCs define two possible equilibrium strategies for agent i, yielding two different expected payoffs. In equilibrium, agent i learns about others only if the gap between these two expected payoffs $\Delta(u)$ more than compensate the fixed cost κ. Since P^* is informationally simple by assumption, we know that this gap is lower than κ. It is hence also lower than κ', and P^* remains a equilibrium under κ': $u \in \mathcal{U}_{IS}(\kappa')$, for all $u \in \mathcal{U}_{IS}(\kappa)$.

Finally, we show that ρ is continuous in κ. By contradiction, suppose it is not: there exists κ^* and $\delta > 0$ such that, for all $\varepsilon > 0$, either $\rho(\kappa^*)-\rho(\kappa^*-\varepsilon) > \delta$ or $\rho(\kappa^*+\varepsilon)-\rho(\kappa^*) > \delta$. Consider the latter case[30] — the function ρ discontinuously jumps up at κ^* — and pick any $\varepsilon > 0$. By assumption there is a difference of at least δ between the Lebesgue measure of $\mathcal{U}_{IS}(\kappa^*+\varepsilon)$ and that of $\mathcal{U}_{IS}(\kappa^*)$. Consider any $u \in \mathcal{U}_{IS}(\kappa^*+\varepsilon) \setminus \mathcal{U}_{IS}(\kappa^*)$. For these utility functions, there exists a non-separable IS equilibrium P^* under $\kappa^*+\varepsilon$ but not under κ^*. Hence $\kappa^* < \Delta(u) < \kappa^*+\varepsilon$: for at least one agent i, it is worth learning about others given that they play P^*_{-i} if the associated fixed cost is κ^* but not if it is $\kappa^*+\varepsilon$. As ε tends to zero, this means that any $u \in \mathcal{U}_{IS}(\kappa^*+\varepsilon) \setminus \mathcal{U}_{IS}(\kappa^*)$ must satisfy $\Delta(u) = \kappa^*$. This equality defines a manifold of dimension strictly less than $|\Omega_i|$ in the domain of i's preferences, and hence the Lebesgue measure of the set of utility functions satisfying it is zero. This contradicts the assumption that the measure of $\mathcal{U}_{IS}(\kappa^*+\varepsilon) \setminus \mathcal{U}_{IS}(\kappa^*)$ must be above δ even for vanishing ε. □

Proof of Proposition 2. Let P^* be an equilibrium in dominant strategy of Γ. That means P_i^* is an optimal strategy for agent i, irrespective of other agents' strategy:

$$P_i^* \in \arg\max_{P_i} \sum_{\omega} \mu_0(\omega)\mathbb{E}_{P_{-i}(\cdot|\omega)}[x_i(m_i,m_{-i})u_i(\omega_i)-t_i(m_i,m_{-i})]-c(P_i) \quad \forall P_{-i}.$$

In particular, P_i^* is optimal when others' strategy is independent of the state,

[30] The proof is similar for the other case.

i.e. when $P_{-i}(\cdot|\omega) = P_{-i}(\cdot|\omega')$ for all ω, ω'. This requires

$$P_i^* \in \arg\max_{P_i} \sum_{\omega} \mu_0(\omega) \left(\mathbb{E}_{P_{-i}(\cdot)}[x_i(m_i, m_{-i})] u_i(\omega_i) - \mathbb{E}_{P_{-i}(\cdot)}[t_i(m_i, m_{-i})] \right) - c(P_i).$$

Note however that in such case, the value of reporting a particular message m_i is state ω equals $\mathbb{E}_{P_{-i}(\cdot)}[x_i(m_i, m_{-i})] u_i(\omega_i) - \mathbb{E}_{P_{-i}(\cdot)}[t_i(m_i, m_{-i})]$, and is always independent of ω_{-i}. Hence, by Assumption 4, agent i's optimal choice rule does not depend on the payoff-irrelevant dimensions ω_{-i}: P_i^* is informationally simple.

The same argument holds for all agents i, and so if P^* is an equilibrium in dominant strategy of Γ then P^* is informationally simple. □

Proof of Theorem 2. The proof of Theorem 2 uses the same techniques as that of Theorem 1. We find necessary conditions for full surplus extraction that are non generic in the space of preferences. By assumption, there is at least one agent i for whom it is efficient to acquire some information. From now on, we restrict attention to this agent i, and take as given that all others play their efficient strategy $P_{-i}^\dagger$. We show that it is generically impossible to induce i to choose her efficient strategy $P_i^\dagger$ while extracting all surplus from her.

Agent i's efficient strategy $P_i^\dagger$ solves

$$\max_{P_i : \Omega \to \Delta\Delta\Omega} \sum_{\omega \in \Omega} \mu_0(\omega) \sum_{\mu_i \in \operatorname{supp} P_i} P_i(\mu_i|\omega) \mathbb{E}_{P_{-i}^\dagger(\cdot|\omega)} \left[\sum_{j \in N} x_j^*(\mu) u_j(\omega_j) \right] - c(P_i).$$

The FOC with respect to $P_i(\mu_i|\omega)\mu_0(\omega)$ writes

$$\mathbb{E}_{P_{-i}^\dagger(\cdot|\omega)} \left[\sum_{j \in N} x_j^*(\mu_i, \mu_{-i}) u_j(\omega_j) \right] - \frac{\partial c(P_i)}{\partial P_i(\mu_i|\omega)\mu_0(\omega)} + \frac{\zeta_i(\omega)}{\mu_0(\omega)} = 0, \tag{1}$$

for all $\mu_i \in \operatorname{supp} P_i$ and for all $\omega \in \Omega$, where $\zeta_i(\omega)$ is the Lagrange multiplier associated with the constraint that $P_i(\cdot|\omega)$ sums to one. Hence agent i's efficient stochastic choice rule $P_i^\dagger$ must satisfy the above system of equations, as well as

the constraints that

$$\sum_{\mu_i} P_i(\mu_i|\omega) - 1 = 0 \quad \forall \omega. \tag{1'}$$

Conversely, the FOCs for the individual decision problem write

$$\mathbb{E}_{P^\dagger_{-i}(\cdot|\omega)}\Big[x_i^*(\mu)u_i(\omega_i) - t_i(\mu_i,\mu_{-i})\Big] - \frac{\partial c(P_i)}{\partial P_i(\mu_i|\omega)\mu_0(\omega)} + \frac{\gamma_i(\omega)}{\mu_0(\omega)} = 0. \tag{2}$$

Surplus extraction requires that the information strategy chosen by the agent coincides with $P_i^\dagger$. Hence $P_i^\dagger$ must solve both (1) and (2). Subtracting (2) from (1) yields:

$$\sum_{\mu_{-i}} P^\dagger_{-i}(\mu_{-i}|\omega)t_i(\mu_i,\mu_{-i}) = -\sum_{\mu_{-i}} P^\dagger_{-i}(\mu_{-i}|\omega)\sum_{j\neq i} x_j^*(\mu)u_j(\omega_j) - \frac{\zeta_i(\omega)-\gamma_i(\omega)}{\mu_0(\omega)}$$

which implies:

$$\mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[t_i(\mu_i,\mu_{-i})\right] = -\mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[\sum_{j\neq i} x_j^*(\mu)u_j(\omega_j)\right] - \sum_{\omega\in\Omega}\big(\zeta_i(\omega)-\gamma_i(\omega)\big). \tag{3}$$

Hence efficient information acquisition requires this particular VCG mechanism. Since, with these transfers, the solutions to both systems of FOCs coincide and equal $P_i^\dagger$, the Lagrange multipliers also coincide: $\zeta_i(\omega) = \gamma_i(\omega)$ for all ω. To extract full surplus from agent i, her expected transfer must sum to her net utility:

$$\mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}[t_i(\mu_i,\mu_{-i})] = \mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[x_i^*(\mu)u_i(\omega_i)\right] - c(P_i^\dagger). \tag{4}$$

Transfers must extract the agent's expected utility given her type while compensating her for the ex-ante investment in information acquisition. Not compensating for these costs would violate the ex-ante IR constraint. Combining (3) and (4), $(P_i^\dagger)_i$ must solve:

$$\mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[\sum_{j\in N} x_j^*(\mu)u_j(\omega_j)\right] = c(P_i^\dagger)$$

Finally taking expectations over μ_i and ω in equation (1) yields:

$$\mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[\sum_{j\in N} x_j^*(\mu)u_j(\omega_j)\right] = \mathbb{E}_{\mu_0,P^\dagger(\cdot|\omega)}\left[\frac{\partial c(P_i^\dagger)}{\partial P_i^\dagger(\mu_i|\omega)\mu_0(\omega)}\right] - \sum_{\omega\in\Omega}\zeta_i(\omega).$$

Combining the above two equations entails that $P_i^\dagger$ must solve:

$$\mathbb{E}_{\mu_0,P_i^\dagger(\cdot|\omega)}\left[\frac{\partial c(P_i^\dagger)}{\partial P_i^\dagger(\mu_i|\omega)\mu_0(\omega)}\right] - \sum_{\omega\in\Omega}\zeta_i(\omega) - c(P_i^\dagger) = 0. \tag{5}$$

That is, they together require that the total cost of information equals the expected marginal cost at the efficient solution. We show that this condition, however, is non-generic.

Define $\hat{\Phi}$ the functional which maps a choice rule for i, Lagrange multipliers and preferences to the LHS of the system of equations (1) and constraints (1′), as well as to the LHS of equation (5). Hence the necessary conditions (1), (1′) and (5) for full extraction of agent i's surplus are jointly written as $\hat{\Phi}(P_i, \zeta_i; u) = \mathbf{0}$. As in the proof of Theorem 1, we leverage the Transversality theorem and Regular Value theorem to show that the set $\hat{\Phi}^{-1}(\mathbf{0}; u)$ is empty for almost all $u \in \mathcal{U}$.

In order to apply the Transversality theorem we need to show that $\mathbf{0}$ is a regular value of $\hat{\Phi}$, i.e. the number of infinitesimally linearly independent equations of the system evaluated at an equilibrium point is maximal:

$$\hat{\Phi}(P_i, \zeta_i; u) = \mathbf{0} \implies \operatorname{rank} D\hat{\Phi}(P_i, \zeta_i; u) = 1 + |\Omega| \times (|\operatorname{supp} P_i^\dagger| + 1).$$

where $D\hat{\Phi}(P, \zeta; u)$ is the Jacobian, and has as many rows as there are equations in the systems (1), (1′) and (5). We need to show that all its rows are linearly independent. The Jacobian has $|\Omega| \times (|\operatorname{supp} P_i^\dagger| + 1) + \sum_j |\Omega_j|$ columns, each corresponding to the derivative w.r.t. each element of $(P_i, \zeta_i; u)$. Ignoring the row that corresponds to equation (5) for now, it equals

$$\begin{array}{cccccccc} \partial P_i & \partial\zeta_i(\omega) & \partial\zeta_i(\omega') & \dots & \partial u_i(\omega) & \partial u_j(\omega) & \dots & \partial u_i(\omega') \end{array}$$

$$
\begin{array}{c}
(1)_{\omega,\mu_i} \\
(1)_{\omega,\mu_i'} \\
\vdots \\
(1)_{\omega',\mu_i} \\
(1)_{\omega',\mu_i'} \\
\vdots \\
(1')_{\omega} \\
(1')_{\omega'} \\
\vdots
\end{array}
\begin{bmatrix}
 & & & \frac{1}{\mu_0(\omega)} & 0 & \dots & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega)}(x_i(\mu_i,\mu_{-i})) & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega)}(x_j(\mu_i,\mu_{-i})) & \dots & \\
 & & & \frac{1}{\mu_0(\omega)} & 0 & & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega)}(x_i(\mu_i',\mu_{-i})) & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega)}(x_j(\mu_i',\mu_{-i})) & & \\
 & -H & & \vdots & & & \vdots & & \ddots & \\
 & & & 0 & \frac{1}{\mu_0(\omega')} & \dots & 0 & 0 & & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega')}(x_i(\mu_i,\mu_{-i})) \\
 & & & 0 & \frac{1}{\mu_0(\omega')} & & 0 & 0 & & \mathbb{E}_{P^\dagger_{-i}(\cdot|\omega')}(x_i(\mu_i',\mu_{-i})) \\
 & & & \vdots & & & \vdots & & & \vdots \\
\mathbf{1} & \mathbf{0} & \dots & 0 & \dots & & & & & \\
\mathbf{0} & \mathbf{1} & & 0 & \dots & & & & & \\
\vdots & & \ddots & & & & & & &
\end{bmatrix}
$$

The columns to the left, which correspond to derivatives w.r.t. P_i, equal the Hessian H of the cost of information as P_i only enters agent i's FOCs through the marginal cost:

$$H = \left[\frac{\partial^2 c(P_i)}{\partial P_i(\mu_i|\omega)\mu_0(\omega)\partial P_i(\mu_i'|\omega')\mu_0(\omega')}\right]_{((\mu_i,\omega),(\mu_i',\omega'))}.$$

Note that columns corresponding to the derivatives w.r.t. $u_j(\omega)$ for some agent j and state ω equal the probability that j gets the good in state ω given $P^\dagger_{-i}$, for each possible report of agent i. It follows directly from Blackwell's principle of irrelevant information that the rows of the above matrix are linearly independent: the efficient allocation must vary with i's report if it is efficient for i to acquire some information.

The key element to prove is that the full surplus extraction condition (5) imposes additional restrictions on $P_i^\dagger$. That is, we need to prove that the derivative of the LHS of (5) w.r.t. (P_i, ζ_i, u) is linearly independent from the rows in the above matrix. The derivative of the LHS of (5) w.r.t. $P_i(\mu_i|\omega)$, $\zeta_i(\omega)$ and $u_i(\omega)$ equal

$$\sum_{\omega'}\mu_0(\omega')\sum_{\mu_i'}P_i(\mu_i'|\omega')\frac{\partial^2 c(P_i)}{\partial P_i(\mu_i'|\omega)\mu_0(\omega')\partial P_i(\mu_i|\omega)\mu_0(\omega)},\ -1, \text{ and } 0, \text{ respectively.}$$

To replicate the derivative w.r.t. P_i from a linear combination of the above matrix, we would need to sum all rows corresponding to the system of equation

(1), weighting each row $(1)_{\omega,\mu_i}$ by $\mu_0(\omega)P_i^\dagger(\mu_i|\omega)$. However, this linear combination also replicates the columns corresponding to derivatives w.r.t. u only if $\mathbb{E}_{\mu_0,P_i^\dagger,P_{-i}^\dagger}[x_i(\mu)] = 0$ for all i, i.e., only if no agent gets the good with positive probability under the efficient solution. That cannot be true if it is socially efficient for agent i to acquire some information.

Thus the Jacobian of $\hat{\Phi}$ at the efficient solution has full rank, and $\mathbf{0}$ is a regular value of $\hat{\Phi}$. The Transversality theorem states, except for a nullset $\mathcal{U}^0 \subset \mathcal{U}$ of preferences, $\mathbf{0}$ is a regular value of $\hat{\Phi}(\cdot; u)$. Then by the Regular Value theorem, $\hat{\Phi}^{-1}(\mathbf{0}; u)$ is a smooth manifold of dimension

$$\dim \hat{\Phi}^{-1}(\mathbf{0}; u) = |\Omega| \times (|\operatorname{supp} P_i^\dagger| + 1) - \Big(1 + |\Omega| \times (|\operatorname{supp} P_i^\dagger| + 1)\Big) < 0.$$

Therefore we conclude that for a full measure set of preferences $u \in \mathcal{U} \setminus \mathcal{U}^0$, the set of stochastic choice rules for i solving (1), (1′) and (5) is empty. In other words, for all preferences in $\mathcal{U} \setminus \mathcal{U}^0$, it is impossible for the designer to both incentivize agent i to choose the efficient strategy and extract all surplus from i.

We have left to show that the set of preferences $\mathcal{U}^0$ for which full surplus extraction is feasible, is closed. This can be done using the same argument as in the proof of Theorem 1, and we omit the formal proof for the sake of brevity. □

Chapter 2

Expectation Formation, Local Sampling and Belief Traps: A new Perspective on Education Choices[1]

Economists usually assume that students form *correct* beliefs about their strategic environment, *independently* from one another. Sociologists, however, argue that students are embedded in their social environment and obtain information by observing the decisions made by others, leading to mistakes and biases. Indeed, there is ample evidence that agents hold incorrect beliefs that are correlated across their social network. To name a few, Kapor et al. (2020) elicit students' subjective admission chances in a low-income district of Connecticut uncovering important departures from rational expectations. Altmejd et al. (2020) show that older sibling's enrollment in college increases a younger sibling's probability of enrolling in college at all, highlighting the importance of the social network on expectation formation. Neither economists nor sociologists, however, possess a coherent framework for thinking strategic interactions between expectation formation and the social environment.

In this paper, we introduce a concrete model of expectation formation in a

[1]This chapter is joint with Philippe Jehiel. We thank Roland Bénabou, Francis Bloch, Gabriel Carroll, Gabrielle Fack, Renato Gomez, Julien Grenet, Marc Gurgand, Yinghua He, Ronny Razin, Al Roth, Olivier Tercieux as well as seminar participants at PSE for useful comments. S. Gleyze acknowledges the support of the EUR grant ANR-17-EURE-0001.

career choice problem. Students differ on two dimensions: their ability which induce different returns to schooling, and their cost of being rejected from elite colleges. We consider a rejection cost instead of an application cost or tuition fees given that fee waivers are now common for disadvantaged students, hence we believe the main obstacle is the social or psychological cost of being rejected from elite institutions. Students strategically choose one out of two occupations: unqualified jobs on the labor market (or non-selective vocational training), and elite colleges. Elite colleges have limited seats and select only the best students up to their capacity. Importantly, we assume that students *do not know* the distribution of admissions conditional on applications.[2] We consider instead that they form their expectations by non-parametrically estimating the distribution of outcomes conditional on a strategy profile, using past experiences from their peers. This estimation is constrained in two ways: First, sample size is endogenous and must be sufficiently large to yield precise estimates. This contrasts with the literature on role models where students learn from a small sample of individuals (Chung, 2000; Bettinger and Long, 2005). Second, students ask in priority peers with similar ability.[3] The rationale is that students know that the admission probabilities depend on ability—though they do not know precisely *how*—, therefore they want to limit biases in their estimates. We introduce the "local sampling equilibrium" in which students best respond to their subjective beliefs, and subjective beliefs are consistent with the above sequential estimation procedure.

Under rational expectations, students perfectly sort in each occupation based on their ability and the equilibrium is efficient. Our main result is that in a local sampling equilibrium two types of inefficiencies arise: First, some high-achieving disadvantaged students self-select out of elite colleges. Second, some average-ability advantaged students apply to elite colleges but are rejected. This equilibrium mismatch is due to the fact that average-students induce a *strategic externality* on high-achieving students by distorting their perceived ad-

[2]This contrasts with the literature on social learning in which agents have prior beliefs about this distribution. We discuss this further in the literature review section.

[3]Our main results hold if students ask random peers, but we believe this is empirically less plausible. Moreover, asking peers with similar ability induce less bias because admission probabilities monotonically increase with ability—hence this assumption is more conservative.

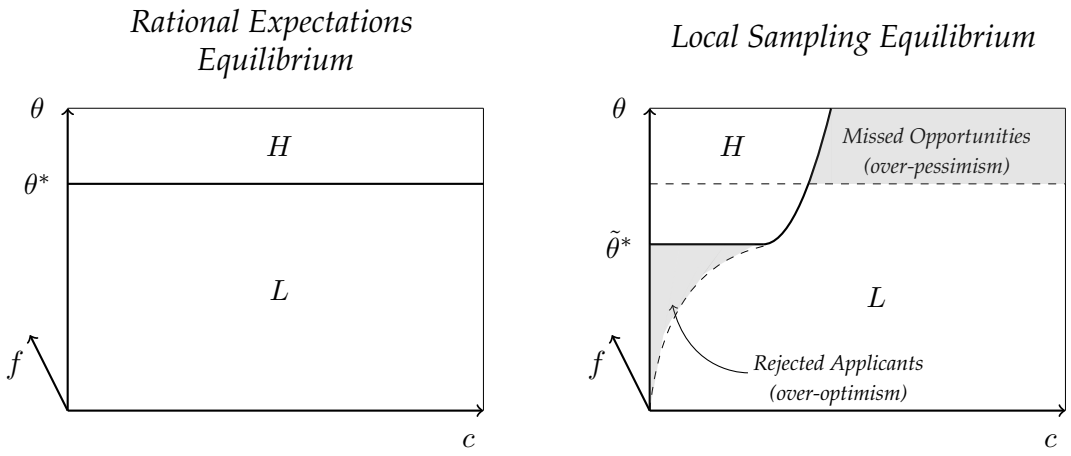


Figure 2.1: *The x-axis represents students' cost, the y-axis represents ability, and the z-axis is the population density. There are two occupations: H are elite colleges that have limited capacity, and L are jobs with no qualifications. (Left) Allocation of students to occupations in a rational expectations equilibrium. (Right) Allocation of students to occupations in a local sampling equilibrium. The shaded areas represent students who are mismatched: the top-right square corresponds to high-achieving disadvantaged students who self-select in non-selective colleges; the bottom-left triangle corresponds to average-achieving advantaged students who apply to elite colleges but are rejected.*

mission chances downward, and on low-achieving students by distorting their beliefs upward. This strategic externality arises due to rationing at elite colleges, hence sample sizes are limited and students must ask average-ability peers to compute their admission chances. Conversely, there is no rationing on the labor market, hence there are no payoff-relevant distortions for students. See Figure 2.1 for a graphical representation of the rational expectations equilibrium and the local sampling equilibrium.

We then investigate the impact of competition across neighborhoods on welfare. First, we show that when a neighborhood becomes relatively wealthier this has a negative impact on self-selection in *other* neighborhoods. This type of cross-neighborhood externality arise because rationing at elite colleges acts as a propagation mechanism of local demand shocks. Second, we investigate the effect of quotas on redistribution and welfare. We find that quotas reduce belief distortion, leading to a better pool of applicants. Quotas act as an effective redistribution tool and increase aggregate welfare. This is in stark contrast

with the "mismatch hypothesis" which asserts that affirmative action necessarily results in minority students being admitted to colleges for which they are otherwise unqualified and reduce welfare. Our result provides a tentative explanation as to why researcher have found no empirical basis for the mismatch hypothesis (Alon and Tienda, 2005; Rothstein and Yoon, 2008; Bertrand et al., 2010). Finally, we investigate the effect of mixed neighborhoods, i.e. reallocating students from poor neighborhoods to rich neighborhoods. When capacity at elite colleges is small, this policy instrument can also increase aggregate welfare.

Our contribution is twofold. At the economical level, we show that equilibrium mismatch arises endogenously when considering a concrete model of expectation formation. This contrasts with the Beckerian model of endogenous schooling with rational expectations that is widespread in empirical and theoretical work. The intuition behind our result is that average-ability students create a strategic externality on high-ability students by distorting their perceived admission chances.

At the methodological level, we introduce a model of expectation formation and show that equilibrium beliefs typically differ from rational expectations. We aim at achieving a balance between strategic sophistication—which is empirically established in college admission (Agarwal and Somaini, 2018)—and the embeddedness of students' beliefs in their social environment. This is in contrast with the "undersocialized" view of an atomic agent that form correct beliefs independently from her environment. Conversely, our model avoids "oversocialized" accounts of expectation formation in which students mechanically inherits the beliefs of their parents.[4]

Related Literature There is a growing literature on expectation formation in education, broadly divided between beliefs on the returns to schooling and subjective admission chances.

The empirical literature on the perceived returns to schooling is mixed. In Wisconsin, Dominitz and Manski (1994) find that the perceived returns from

[4]The distinction between undersocialized and oversocialized explanations is due to Granovetter (1985).

a Bachelor's degree compared to a high school diploma are positive. In Chile, Hastings et al. (2015) show that low-achieving disadvantaged students who apply to low-earning college degree programs overestimate earnings for past graduates by over 100%, while beliefs for high-achieving students are correctly centered. Conversely in the Dominican Republic, Jensen (2010) find that the perceived returns to secondary school are extremely low, despite high measured returns.

Very few papers investigate subjective admission chances, which is the focus of our paper. Most notably, Hastings and Weinstein (2008) show that providing information about school quality and odds of admission to low-income families with high-achieving students increases application to good schools. It is unclear, however, if the effect is driven by growing awareness about these schools or changing expectations. Kapor et al. (2020) directly elicit admission probabilities of students facing a centralized school choice mechanism that rewards strategic behavior. They find that households play strategically, but do so with miscalibrated beliefs. Belief errors, however, do not seem to correlate with observable characteristics such as race or economic status. Finally, Altmejd et al. (2020) show that older sibling's enrollment in a better college increases a younger sibling's probability of enrolling in college at all, especially for families with low predicted probabilities of enrollment.

The first theoretical model of expectation formation on the returns to schooling is due to Manski (1993). He postulates an additive log-income equation, and he assumes that students infer the returns to schooling by taking the conditional expectation of log-income. If students omit to condition on ability—e.g., because they do not observe the ability of their peers—he shows that more low-ability and less high-ability students enroll to college.

There is a vast literature on social learning illustrating that past cohorts' behavior influences the expectations of current cohorts (Banerjee, 1992; Bikhchandani et al., 1992; Ellison and Fudenberg, 1995). These papers, however, typically assume that agents have enough prior information to infer the outcome of counterfactual actions using Bayes' rule. Manski (2004) relaxes this assumption by considering a social learning environment in which students have no prior belief on the distribution of outcomes conditional on actions—as in our model.

Hence students cannot infer anything on counterfactual actions. Only assuming the stationarity of the outcome distribution—as we do in this paper[5]—he shows that learning induces a process of sequential reduction in ambiguity. Though similar in motivation, our papers differ with the social learning literature because we account for strategic interactions among students which are instrumental to produce belief distortion.

Finally, several papers on bounded rationality in games introduce more realistic models of expectation formation. For instance, Jehiel (2005) introduces a model of coarse expectations in which players bundle actions into classes. In equilibrium, players best-respond to their analogy-based expectations, and expectations correctly represent the average behavior in every class. Our paper introduces a different learning rule where students average the outcome of an endogenously chosen group of players and do not bundle actions, whereas in Jehiel (2005) players average the outcome of an exogenously given bundle of actions using past observations from an exogenously given group of players. He justifies belief consistency, as we do here, using a learning argument (Fudenberg and Levine, 1998).

2.1 Setup

We introduce a stylized model of career choice with strategic students and rationing at elite colleges. There is a unit mass of students indexed by their ability $\theta \in [\underline{\theta}, \overline{\theta}] \subseteq \mathbb{R}_+$, and by their cost $c \in [\underline{c}, \overline{c}] \subseteq \mathbb{R}_+$. There is a probability distribution F on $N \equiv [\underline{\theta}, \overline{\theta}] \times [\underline{c}, \overline{c}]$ with continuous density f that has full support. We will consider two types of costs: either an opportunity cost from being rejected from elite colleges which captures higher marginal utility of money (i.e., conditional on being rejected at an elite college, poorer students would have benefited more from going directly on the labor market), or an application cost. We will do most of the analysis in the main text using the opportunity cost, and explain how it differs from the application cost in the Appendix.

Students choose among two occupations: going directly on the labor market (or a non-selective vocational training) L, or applying to selective colleges H.

[5]Meaning that colleges never modify their admission criteria.

Without loss of generality, the utility of attending an elite college is $U^H(\theta) = \theta$, whereas for simplicity we assume that the utility of going directly on the labor market is $U^L(\theta) = 0$ for all θ.

Students can apply to only one occupation: the action space is then $A = \{L, H\}$. There is no rationing for going on the labor market. Elite colleges, however, have a limited number of seats and they select students with the highest ability (among the pool of applicants) up to their capacity $q \ll 1$.[6] The payoffs are as follows:

- If student (θ, c) goes on the labor market L her utility is 0.
- If student (θ, c) applies to H and obtain a seat, her utility is θ.
- If student (θ, c) applies to H but does not get a seat, she goes on the labor market and her utility is $-c$.

A strategy profile $\sigma : N \longrightarrow \Delta A$ is a (mesurable) function from the population of students to mixed actions. This is a binary action game, hence we let $\sigma(\theta, c) \in [0, 1]$ simply denote the probability that student (θ, c) applies to H.

A key object that drives the choice of student (θ, c) is the subjective probability this student (subjectively) assigns to obtaining a seat at an elite college conditional on applying to H. In both the rational case and our approach, this subjective probability turns out to depend only on θ and we denote it by $p(\theta)$ accordingly. Based on $p(\theta)$, student (θ, c) applies to H whenever

$$p(\theta)\theta - (1 - p(\theta))c \geq 0$$

This leads to the following definition of an optimal strategy profile.

DEFINITION 2. *σ is optimal given subjective beliefs $p(\cdot)$ if*

$$\sigma(\theta, c) = \begin{cases} 1 & \textit{when} \quad c \leq \frac{p(\theta)}{1-p(\theta)}\theta \\ 0 & \textit{when} \quad c > \frac{p(\theta)}{1-p(\theta)}\theta \end{cases}$$

[6] Our results are unchanged if colleges only receive a noisy signal about students' ability.

For any strategy profile, let $\theta(\sigma)$ denote the cutoff at H such that any student with ability $\theta > \theta(\sigma)$ who applies to H is admitted. It is defined as follows: $\theta(\sigma) = \underline{\theta}$ when

$$\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{c}}^{\overline{c}} \sigma(\theta, c) f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta < q$$

Otherwise, $\theta(\sigma)$ is uniquely defined as the largest θ^* such that

$$\int_{\theta^*}^{\overline{\theta}} \int_{\underline{c}}^{\overline{c}} \sigma(\theta, c) f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta = q$$

Subjective beliefs are rational when they are consistent with the admission cutoff, given a strategy profile.

DEFINITION 3. *$p^R(\cdot)$ is rationally consistent with σ if*

$$p^R(\theta) = \begin{cases} 1 & \textit{when} \quad \theta \geq \theta(\sigma) \\ 0 & \textit{when} \quad \theta < \theta(\sigma) \end{cases}$$

Therefore, the rational expectations equilibrium is defined as follows.

DEFINITION 4 (Rational Expectations Equilibrium)**.** *σ^R is a rational expectations equilibrium if there exist subjective beliefs p^R such that σ^R is optimal given p^R and p^R is rationally consitent with σ^R.*

Let us now characterize the unique rational expectations equilibrium—thus proving existence. Given the strategy profile σ, define the admission cutoff $\theta^* = \theta(\sigma)$ as the highest θ such that a mass q of students are admitted to H. Subjective beliefs are consistent with σ, hence it is optimal to apply to H for all students with ability $\theta > \theta^*$. Assuming independence between ability and cost $F(\theta, c) = H(\theta)G(c)$, the admission cutoff θ^* solves

$$\int_{\theta^*}^{\overline{\theta}} h(\theta) \, \mathrm{d}\theta = q \iff \theta^* = H^{-1}(1 - q). \tag{2.1}$$

Therefore, the equilibrium allocation of students to occupations can be described with a unique cutoff $H^{-1}(1 - q)$.

PROPOSITION 3 (Equilibrium Characterization). *Assume that ability and cost are independent. In the unique rational expectations equilibrium, students* $N^H = \{(\theta, c) : \theta > H^{-1}(1-q)\}$ *obtain a seat at elite colleges, and* $N^L = N \setminus N^H$ *go on the labor market.*

(All formal proofs and verification arguments are deferred to the Appendix). The rational expectations equilibrium induces perfect assortative matching as students sort across occupations based on their ability. Namely, high-achieving students go to elite colleges, and average- or low-ability students go on the labor market. See Figure 2.1 (Left) above for a graphical illustration of the equilibrium.

Define welfare as

$$W(\sigma) = \int_{\theta^*}^{\overline{\theta}} \int_0^{c^H} \theta f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta - \int_0^{\theta^*} \int_0^{c^H} c f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta$$

where $c^H(\theta, p(\theta))$ is the cost below which student (θ, c) applies to H conditional on admission chances p. In the rational expectation equilibrium, $c^H(\theta, p(\theta)) = \overline{c}$ for all $\theta > H^{-1}(1-q)$ and $c^H(\theta, p(\theta)) = 0$ for all $\theta > H^{-1}(1-q)$. Rational expectations induce perfect sorting which is welfare maximizing.

2.2 Expectation Formation and Belief Traps

In this section we introduce a concrete model of expectation formation, and we show that it leads to persistent belief distortions among high-achieving disadvantaged students—so-called "belief traps."

Students have no prior over the distribution of admissions conditional on applications. We assume that they non-parametrically estimate this distribution by averaging the outcome of their peers who are closest to them in terms of ability. Let $\mathcal{B}(N)$ denote the set of measurable subsets of N.

DEFINITION 5. *The sample for action* H *of student* (θ, c) *conditional on a strategy profile* σ *(from the previous generation) is*

$$S(\theta, c \mid \sigma) = \operatorname*{arg\,inf}_{B \in \mathcal{B}(N)} \left\{ \int_B |\theta - \tilde{\theta}| \, \mathrm{d}F(\tilde{\theta}, \tilde{c}) : \int_B \sigma(\tilde{\theta}, \tilde{c}) \, \mathrm{d}F(\tilde{\theta}, \tilde{c}) > \tau \right\}.$$

If a mass less than τ plays action H, we set $S(\theta, c \mid \sigma) = N$ for definiteness. In words, S is the mass τ set of students with ability closest to θ. There is a convex penalty of including students with dissimilar ability, hence the sample $S(\theta, c \mid \sigma)$ is rectangular and it can be described by a simple index:

$$b(\theta, \sigma) = \inf \left\{ b > 0 : \int_{\max\{\underline{\theta}, \theta - b\}}^{\min\{\overline{\theta}, \theta + b\}} \int_{\underline{c}}^{\overline{c}} \sigma(\tilde{\theta}, \tilde{c}) \, \mathrm{d}F(\tilde{\theta}, \tilde{c}) > \tau \right\}.$$

This means that the sample for action H of student (θ, c) is obtained by taking all students with ability $\theta' \in [\theta - b(\theta, \sigma), \theta + b(\theta, \sigma)]$ regardless of their cost. See Figure 2.2 below for a graphical illustration.

Remark 1. Other sampling rules, such as sampling uniformly at random or sampling students with similar costs, would not change qualitatively the results. It would only increase belief distortions because the true admission chances depend on ability only. Therefore, our results can be thought of a lower bound on belief distortions.

We can now define subjective admission chances. As in the previous section, we denote $\theta(\sigma)$ the admission cutoff at elite colleges given the strategy profile σ. The subjective admission chances at elite colleges H are obtained by averaging the experiences of the students in the sample.

DEFINITION 6. *Subjective admission chances at elite colleges p are τ-consistent with σ if*[7]

$$p(\theta) = \frac{1}{\tau} \int_{S(\theta, c \mid \sigma)} \sigma(\tilde{\theta}, \tilde{c}) \mathbf{1}\{\tilde{\theta} > \theta(\sigma)\} \, \mathrm{d}F(\tilde{\theta}, \tilde{c}).$$

We now introduce our solution concept, the local sampling equilibrium, which requires optimality of actions and consistency of beliefs.

DEFINITION 7 (Local Sampling Equilibrium)**.** *σ is a local sampling equilibrium if there exists p such that σ is optimal given p and p is τ-consistent with σ.*

[7] If a mass of students less than τ chooses H, then we divide by $\int_B \sigma^k(\tilde{\theta}, \tilde{c}) \, \mathrm{d}F(\tilde{\theta}, \tilde{c})$ instead of τ.

We interpret this solution concept as the stationary point of an intergenerational model of learning in which students of the current generation ask peers from the previous generation the outcome of their behavior. Therefore, this sample is completely endogenous as it depends on the strategy profile of the previous generation. Importantly, students know nothing ex-ante about the admission process: it could be either because schools do not disclose their admission criteria, or because students lack the ability to understand the admission process, or because they do not trust publicly disclosed information. Therefore, students entirely rely on the information provided by their social network. Of course, this is a stylized assumption and in practice we expect students to use a mix of information sources to form their expectations.

We made two assumptions on the learning process. First, students care about the precision of their estimate hence they must acquire a sufficient amount of data for each action. Formally, this means that students ask a mass $\tau \in (0, 1]$ of students from the previous generation, where τ is interpreted as a confidence parameter. This parameter captures a bias-variance trade-off: if the sample it too small then subjective admission chances are unbiased because they are computed using students with similar ability, but the estimator is noisy.[8] Conversely, if the sample is too large then subjective admission chances are precisely estimated but they are more likely to be biased.

Second, students contact in priority peers with similar ability. There are two possible justifications. From a statistical perspective, if students know that the admission probability is somewhat correlated with their ability, then they might reduce bias by asking peers with similar ability.[9] From a sociological perspective, if students have homophilic preferences their close ties are more likely to have similar ability.

Note that students include in their sample for action k only peers who *actually* played action k in the previous period. Therefore, students make no

[8] This is a reduced-form interpretation because there is no actual noise in the estimate as students sample from a continuum of peers.

[9] Assuming that students contact peers with similar cost or assuming random contact would induce more bias (because the admission probability only depends on ability) hence it would only strengthen our result. Conservatively, we assume that students contact peers with similar ability to limit biases. Therefore, our results can be thought of a lower bound on the bias.

inference using counterfactual outcomes—i.e. they are not asking their peers "What would have been your admission chances at x conditional on applying there?". Who is included in the sample is endogenous and typically differ for each player and for each action, even though sample size is identically equal to τ for each player and for each action. Concretely, the perimeter of the sample for H of low-ability disadvantaged students is very large because no close ties ever apply to H. Therefore, they will need to ask high-achieving peers that have very different characteristics which induce a large bias in the subjective admission chances. In general, a large perimeter is synonym of a larger bias because the sample includes students with very different characteristics, whereas a small perimeter is synonym of a smaller bias.

Existence We apply a fixed point argument on the map from subjective beliefs $p : \Theta \longrightarrow [0, 1]$, to best responses σ, to subjective beliefs computed from the best response. The fixed point exists if each sub-map is continuous. It is easy to see that the best response σ has a threshold structure that varies continuously with p. Moreover, the sample bounds $b(\theta, \sigma)$ are continuous in the strategy profile σ, and so are subjective beliefs p. This shows the existence of a pure strategy local sampling equilibrium. (The formal proof is deferred to the Appendix).

Equilibrium Characterization Fixing ability and the subjective admission chances, students who apply to H have a cost lower than

$$c^H(\theta, p(\theta)) = \frac{p(\theta)}{1 - p(\theta)}\theta.$$

Define the mass of applicants to H as follows:

$$\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{c}}^{c^H(\theta, p(\theta))} f(\theta, c)\, \mathrm{d}c\, \mathrm{d}\theta.$$

In a local sampling equilibrium, the ability of the last student admitted to H, denoted $\tilde{\theta}^*$, is such that the mass of applicants at H is equal to the capacity of

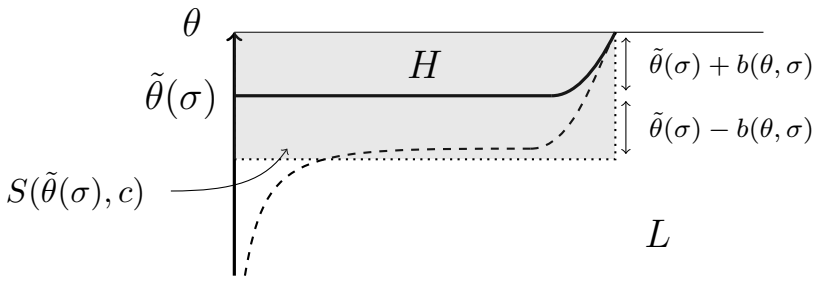


Figure 2.2: *Construction of the sample for the last student admitted at an elite college* $(\tilde{\theta}(\sigma), c)$ *in* (c, θ)*-space. The sample, represented in the shaded box, includes approximately a mass* τ *of students who applied to an elite college* H*. All students above the dashed line applied to* H *(i.e.* $\sigma(\theta, c) = 1$*) but only those above the solid line got admitted at an elite college. Rejected students exert a strategic externality on higher achieving students by distorting their estimated admission chances downard.*

elite colleges:

$$\int_{\tilde{\theta}^*}^{\overline{\theta}} \int_{\underline{c}}^{c^H(\theta, p(\theta))} f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta = q. \tag{2.2}$$

Assuming the independence between ability and cost $F(\theta, c) = H(\theta)G(c)$, we can rewrite this equation as

$$\int_{\tilde{\theta}^*}^{\overline{\theta}} h(\theta) G\left(\frac{p(\theta)}{1 - p(\theta)}\theta\right) \, \mathrm{d}\theta = q \tag{2.3}$$

Let us now derive the equation that guarantees τ-consistency of subjective admission chances. The subjective admission chances of student (θ, c) are τ-consistent if they solve the following equation:

$$p(\theta) = \frac{1}{\tau} \int_{\max\{\underline{\theta}, \theta - b(\theta,\sigma)\}}^{\min\{\overline{\theta}, \theta + b(\theta,\sigma)\}} \int_{\underline{c}}^{c^H(\tilde{\theta}, p(\tilde{\theta}))} \mathbf{1}\{\tilde{\theta} > \tilde{\theta}^*\} \, \mathrm{d}F(\tilde{c}, \tilde{\theta}).$$

Assuming that ability and costs are independent, this writes:

$$p(\theta) = \frac{1}{\tau} \int_{\max\{\underline{\theta}, \theta - b(\theta,\sigma)\}}^{\min\{\overline{\theta}, \theta + b(\theta,\sigma)\}} G\left(\frac{p(\tilde{\theta})}{1 - p(\tilde{\theta})}\tilde{\theta}\right) \mathbf{1}\{\tilde{\theta} > \tilde{\theta}^*\} \, \mathrm{d}H(\tilde{\theta}). \tag{2.4}$$

In equilibrium, $\tilde{\theta}^*$ must solve (2.2) given $p(\theta)$, and $p(\theta)$ must solve (2.4) for all students (θ, c) given $\tilde{\theta}^*$.

We can now compare equation (2.3) with the equation that defines the last student admitted to H in a rational expectations equilibrium:

$$\int_{\theta^*}^{\bar{\theta}} h(\theta)\,\mathrm{d}\theta = q. \tag{2.5}$$

If there are students with sufficiently high costs—e.g. if g has full support on $\mathbb{R}_+$—any small belief distortion in equation (2.4) will induce self-selection among disadvantaged students: $c > c^H(\theta, p(\theta))$. Then, the term under the integral sign in (2.3) is smaller than in (2.5) because $G(c^H(\theta, p(\theta))) < 1$ as $c^H(\theta, p(\theta)) < c \leq \bar{c}$. Therefore, the ability of the last admitted student at H in a local sampling equilibrium $\tilde{\theta}^*$ must be smaller than in a rational expectations equilibrium to fill all the seats in equation (2.3).

We just proved that two types of inefficiencies arise in a local sampling equilibrium: high-achieving disadvantaged students self-select out of elite colleges even though their actual admission probability is one, and low-achieving advantaged students spend inefficient resources in applications at elite colleges even though their actual admission chances are zero. See Figure 2.1 in the introduction for a graphical representation of the two inefficiencies.

PROPOSITION 4 (Equilibrium Characterization). *Suppose that g has full support on $\mathbb{R}_+$ and assume ability and cost are independent. There is a local sampling equilibrium such that students $\tilde{N}^H = \{(\theta, c) : \theta > \tilde{\theta}^*, c \leq c^H(\theta, p(\theta))\}$ obtain a seat at elite colleges and $N^L = N \setminus \tilde{N}^H$ go on the labor market. There are two types of inefficiencies:*

1. *Missed opportunities: all students (θ, c) with ability $\theta > \theta^*$ and cost $c > c^H(\theta, p(\theta))$ self-select out of elite colleges.*

2. *Inefficient applications: all students (θ, c) with ability $\theta < \tilde{\theta}^*$ and cost $c < c^H(\tilde{\theta}, p(\tilde{\theta}))$ apply to H but are rejected and suffer a cost $-c$.*

Observe that compared to the rational expectations equilibrium both the supply side and the demand side suffer from inefficiencies. On the supply side, belief distortion arises endogenously and leads to payoff-relevant mistakes for

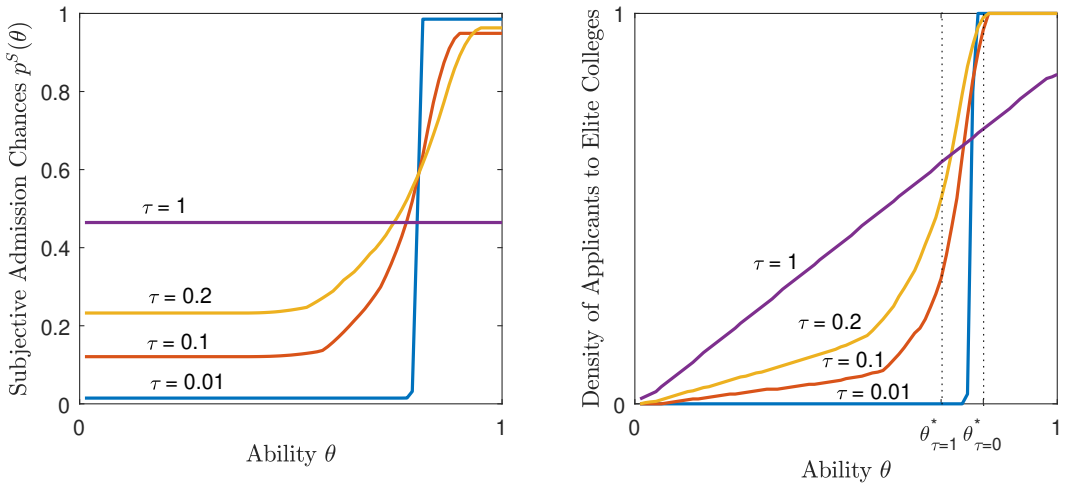


Figure 2.3: *(Left) Subjective admission chances as a function of student ability. Bias in subjective beliefs increases with the confidence parameter τ. (Right) Density of applicants to H as a function of student ability. As τ increases, the admission cutoff $\tilde{\theta}^*$ decreases, the number of self-selecting students (on the right of the cutoff) increases and the number of inefficient applicants (on the left of the cutoff) increases as well.*

high-achieving students and low-achieving advantaged students. On the demand side, the quality of the pool of admitted students at elite colleges is lower than with rational expectations due to equilibrium mismatch.

We now describe comparative statics with respect to the confidence parameter τ. When $\tau \to 0$ students form their expectations using an infinitesimal sample of individuals. As it turns out, in our model this leads to rational expectations because students do not bias their estimate with dissimilar students. Indeed, taking the limit $\tau \to 0$ of the implicit equation (2.4) we can see that if $\theta < \tilde{\theta}^*$ then there is τ_* small enough such that $\theta + b(\theta, \sigma) < \tilde{\theta}^*$ and $\theta - b(\theta, \sigma) < \tilde{\theta}^*$. Therefore, the integral in (2.4) is zero, and we have $p(\theta, c) = 0$. Similarly, one can verify that for all $\theta > \tilde{\theta}^*$, $p(\theta, c) = 1$. Therefore, only the best students apply to elite colleges and the last student admitted in a local sampling equilibrium coincides with that of rational expectations $\tilde{\theta}^* = \theta^*$.

Students, however, do not form expectations using one data point. To reduce risk induced by imprecise estimates, they are more likely to include the outcome of multiple peers. In our model, belief distortions increase with the

confidence level τ because students include peers with very different characteristics in their sample. Hence bias in the estimate stems from a selection bias that increases with τ. As $\tau \to 1$ (i.e., students include the entire population), the subjective beliefs of the entire population converge. In practice, we would expect intermediary values of τ so as to trade-off bias and precision of the estimate.

This comparative statics is illustrated in Figure 2.3 (Left). Figure 2.3 (Right) illustrates the two types of inefficiencies that arise in a local sampling equilibrium. We see that as the confidence parameter τ increases, the admission cutoff $\tilde{\theta}^*_\tau$ decreases. Subjective beliefs, however, move smoothly around this threshold hence the mass of student who apply to H with an ability that is below the cutoff $\tilde{\theta}^*_\tau$ is positive (inefficient applications), and the mass of students who apply to H with an ability that is above the cutoff is below one (missed opportunities).

PROPOSITION 5. *In any local sampling equilibrium, a higher confidence parameter τ leads to more self-selection from high-achieving disadvantaged students and to more inefficient applications from low-achieving advantaged students. Conversely, for $\tau \longrightarrow 0$ the local sampling equilibrium converges to the rational expectations equilibrium.*

We conclude this section with the case in which equilibrium subjective admission chances can be characterized in closed form. This will prove useful when studying competition across neighborhoods in the next section.

2.2.1 Large sampling window $\tau = 1$

When $\tau \to 1$, we can solve efficiently for the equilibrium posterior beliefs by noticing that all students must have identical beliefs (hence p is independent of θ). For a given capacity q, subjective admission chances are equal to the capacity at elite colleges divided by the mass of applicants:

$$p = q \Big/ \int_0^1 G\left(\frac{p}{1-p}\theta\right) \mathrm{d}\theta$$

Given that $p \mapsto p \int_0^1 G(\frac{p}{1-p}\theta)\,\mathrm{d}\theta$ is a strictly increasing function of p with value 0 at $p = 0$ and 1 at $p = 1$, we obtain that for each q there is a unique $p(q)$ satis-

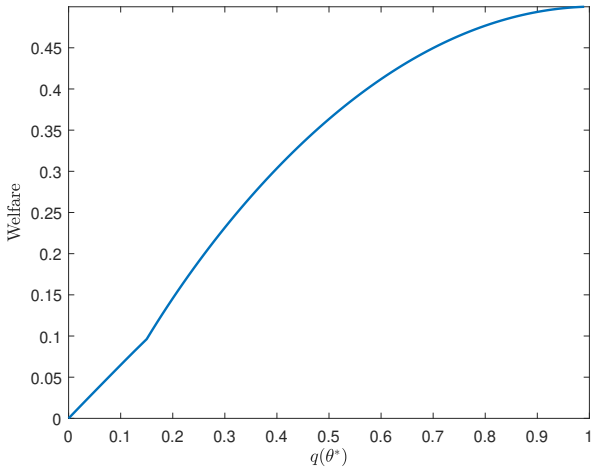


Figure 2.4: *Total welfare is monotonic in the number of allocated seats* $q(\theta^*)$.

fying the above equation. As will be convenient when analyzing the multiple neighborhood case, it is useful to parameterize the equilibrium by the admission threshold θ^* defined for a given q by

$$\int_{\theta^*}^{1} G\left(\frac{p}{1-p}\theta\right) \mathrm{d}\theta = q$$

where p is $p(q)$ as previously defined.

We can now define a function which takes the value 0 at an equilibrium belief p:

$$H(p;\theta^*) = \frac{\int_{\theta^*}^{1} G(\frac{p}{1-p}\theta)\,\mathrm{d}\theta}{\int_{0}^{1} G(\frac{p}{1-p}\theta)\,\mathrm{d}\theta} - p.$$

The first term should be understood as the ratio between the number of accepted students to the number of applicants. Hence, in equilibrium this should be equal to p when the sampling window is $\tau = 1$. To guarantee the existence of a root to the equation $H(p;\theta^*) = 0$, we make the following assumption.

Assumption 5. *The function* $H(p;\theta^*)$ *is decreasing in* p.

This assumption is satisfied when G is a uniform distribution, which we are going to assume in most of the analysis of the next section. Moreover, when G

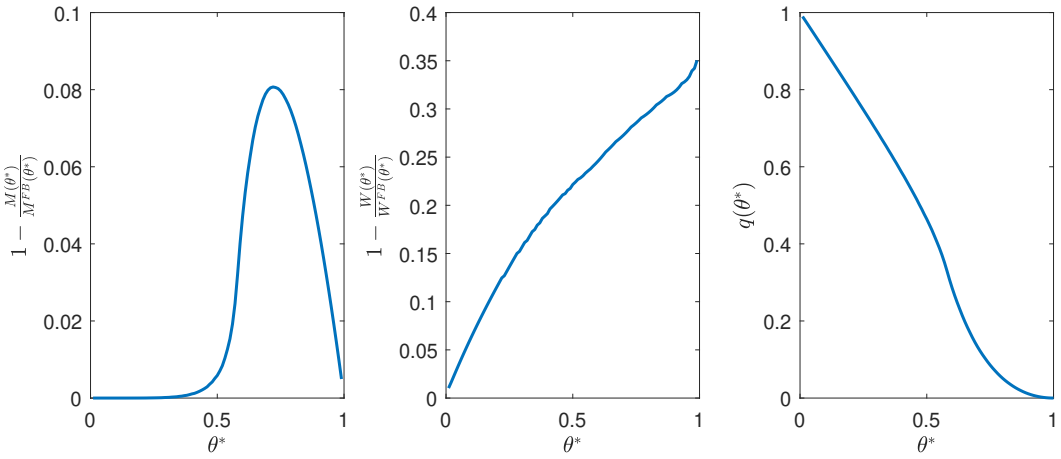


Figure 2.5: *Simulation for $G(\theta) = c/2$. (Left) Quality loss in a local sampling equilibrium as a function of the admission threshold θ^* (Middle) Welfare loss in a local sampling equilibrium as a function of the admission threshold θ^*. (Right) Number of seats allocated in equilibrium as a function of the admission threshold θ^*.*

is a uniform distribution welfare is monotonically increasing in the number of allocated seats, as show in Figure 2.4 (this holds for any support of g).

It is useful to define measures of quality loss and welfare loss between a local sampling equilibrium and the first-best (achieved in a rational expectations equilibrium). Let the average ability (θ) of admitted students be

$$M(\theta^*) = \frac{\int_{\theta^*}^{1} G\left(\frac{p(\theta^*)}{1-p(\theta^*)}\theta\right)\theta \, \mathrm{d}\theta}{\int_{\theta^*}^{1} G\left(\frac{p(\theta^*)}{1-p(\theta^*)}\theta\right) \, \mathrm{d}\theta}$$

to be compared to $M^{FB}(\theta^*) = 1 - \frac{q(\theta^*)}{2}$, the corresponding first-best average quality when there are $q(\theta^*)$ seats. Similarly, $W(\theta^*)$ is defined as in Section 2 and $W^{FB} = q(\theta^*)(1 - \frac{q(\theta^*)}{2})$. In Figure 2.5, we plot the ratio of these quantities to assess the relative loss of quality and welfare induced by biased beliefs.

For a uniform distribution, we can solve for equilibrium beliefs in closed form. Suppose that ability and cost are uniformly distributed: $F(\theta) = \theta$ on

$[0,1]$ and $G(c) = \frac{c}{\overline{c}}$ on $[0,\overline{c}]$. Subjective admission chances solve

$$p = q \Big/ \left[\int_0^1 \min \left\{ \frac{\frac{p}{1-p}\theta}{\overline{c}}, 1 \right\} \mathrm{d}\theta \right]. \tag{2.6}$$

We first consider the case in which the minimum in the previous equation does not bind. Hence, equation (2.6) is a simple quadratic function and solving for p yields:

$$p = -q\overline{c} + \sqrt{q^2\overline{c}^2 + 2q\overline{c}} \tag{2.7}$$

We now characterize under what conditions the minimum does not bind. Substituting the expression for p in the following equation:

$$\frac{\frac{p}{1-p} - \underline{c}}{\overline{c} - \underline{c}}\theta < 1$$

and solving for q yields

$$\hat{q}(\theta) = \frac{\overline{c}^2}{2\theta^2\left[(1+\frac{\underline{c}}{\theta})^2\overline{c} - 2(1+\frac{\underline{c}}{\theta})\frac{\underline{c}}{\theta}\overline{c}\right]}$$

Therefore, the subjective beliefs are given by (2.7) whenever the capacity at elite colleges verifies $q < \sup_\theta \hat{q}(\theta)$. This is the relevant case in practice given the very high level of competition at elite colleges. If this condition is violated, however, we define $\hat{\theta}$ as the ability that solves $q = \hat{q}(\hat{\theta})$ and we decompose the integral in (2.6) into two integrals on the intervals $[0,\hat{\theta}]$ and $[\hat{\theta},1]$ and then solve for p accordingly.

For intermediate values of τ, it is not possible to obtain closed form solutions, therefore we run simulations in Figure 2.3. As we already discussed, the bias increases with the confidence parameter τ because of selection of students with dissimilar characteristics in the sample.

2.3 Competing Neighborhoods

We consider now the case of multiple neighborhoods competing for the same positions. The neighborhood plays a role only in shaping the samples from which students form their subjective assessment, as we assume the sampling is made locally (only within the neighborhood to which the student belongs). The fact that students from the various neighborhoods compete for the same seats creates a linkage between the various neighborhoods as the threshold ability θ^* above which students get admitted has to be the same across neighborhoods. This linkage in turn induces externalities across neighborhoods the effects of which is the main subject of interest of this Section. To formalize the questions of interest, consider a two-neighborhood setup. Neighborhood $i = 1, 2$ consists of a unit mass of students with (θ_i, c_i) distributed according to distribution f_i and sampling window τ_i. Consider first neighborhood i in isolation, assume there is a mass q_i of seats available for students in this neighborhood and that students follow strategy σ_i. We let $\theta(\sigma_i, q_i)$ be the corresponding threshold admission ability in this neighborhood. It is computed as shown in Section 2. An equilibrium is formally defined as follows.

DEFINITION 8. *An equilibrium with competing neighborhoods $i = 1, 2$ (with characteristics f_i and τ_i) and mass q of seats is a strategy profile (σ_1, σ_2) such that there exist q_1, q_2 satisfying*

1. *σ_i is a sampling equilibrium in the neighborhood i with a mass q_i of seats;*
2. *$q_1 + q_2 = q$ and,*
3. *$\theta(\sigma_1, q_1) = \theta(\sigma_2, q_2)$.*

The definition of welfare W_i and average ability of admitted students M_i in neighborhood i are adapted accordingly. Denote $W = W_1 + W_2$ the aggregate welfare, and $M = \frac{q_1 M_1 + q_2 M_2}{q_1 + q_2}$ the average ability of admitted students.

In this section, we are interested in (i) the strategic interactions across neighborhoods, and (ii) how asymmetries across neighborhoods impact welfare and the average quality of admitted students. We consider asymmetries in sampling window, and asymmetries in cost distributions—fixing the distribution

of ability. When varying τ we will consider that τ is either 0 or 1 to make things simpler. When considering asymmetric distributions, we will consider that in both neighborhoods θ_i is uniformly distributed on $(0,1)$ and c_i is distributed according to cdf G_i, independently of θ_i. G_i will be taken to be a uniform distribution on $[\underline{c}_i, \overline{c}_i]$ in most results and simulations.

Finally, we investigate two types of policies aimed at reducing such inequalities: quotas and mixed neighborhoods (i.e., directly changing the composition of neighborhoods).

2.3.1 Asymmetries in Sampling Window

We first investigate asymmetries in sampling windows, namely $\tau_i \neq \tau_j$. This arises naturally when neighborhoods are of different size, and students ask a fixed numbers of peers to construct their estimate. In this case, students in the smaller neighborhoods mechanically communicate with a larger fraction of their peers.

To keep things simple, we consider an extreme situation where the sampling window in neighborhood i goes to zero (i.e., neighborhood i is very large) whereas in j students contact all their peers (i.e., neighborhood j is very small). We show that neighborhood j is disadvantaged and obtain less seats at elite colleges.

PROPOSITION 6. *Suppose that $G_i = G_j$ and consider a sequence (τ_i^n) such that $\tau_i^n \longrightarrow 0$ and $\tau_j = 1$, then $\lim_{n\to\infty} q_i^n > q_j$. If G_i and G_j are uniform, this implies that $\lim_{n\to\infty} W_i^n > W_j$.*

To understand the result, observe that the set of admitted students is identical whether $\tau_i^n \longrightarrow 0$ or $G = \delta_0$ (point mass at zero cost). Indeed, as the sampling window becomes smaller, the sampling bias on p_i goes to zero and each student has an asymptotically unbiased estimator of his admission chances. Therefore, students apply to H if and only if $\theta_i \geq \theta^*$. Instead, when $G = \delta_0$ all students apply to H, and only students with $\theta_i \geq \theta^*$ are admitted. Therefore, the set of admitted students is identical in both cases. Now, when the cost distribution goes to zero, it is quite intuitive that students never self-select and take a larger number of seats at elite colleges.

2.3.2 Asymmetries in Cost Distribution

We now investigate asymmetries in cost distribution, i.e. $G_i \neq G_j$. This can arise due to differences in social norm for instance: the cost of not attending an elite college might be higher in some communities than others.

To keep things simple, we consider an extreme situation where the cost is zero in neighborhood i (i.e. $G_i = \delta_0$) whereas the cost is arbitrary but non-zero in neighborhood j. We show that neighborhood j is disadvantaged and obtain less seats at elite colleges.

PROPOSITION 7. *Suppose that $\tau_i = \tau_j$, and $G_i = \delta_0$ but $G_j \neq \delta_0$. Then, $q_i > q_j$. If G_j is a uniform distribution, this implies that $W_i > W_j$.*

Comparative statics with respect to the cost distribution, however, are not always intuitive in our model. Using simulations, we show that a first order stochastic shift in G_j with respect to G_i does not necessarily imply that W_j/W_i decreases, as one might expect.

We consider two cases. First, a situation in which cost is uniformly distributed on $[0, 0.1]$ in neighborhood i and $[0, 0.3]$ in neighborhood j. Second, a situation in which cost is uniformly distributed on $[0, 0.5]$ in neighborhood i and $[0.5, 1]$ in neighborhood j. (See Figure 2.3.3 below, case without quotas). In the second situation, the "disadvantaged" neighborhood has more seats than the "advantaged" neighborhood. This counter-intuitive effect is due to the fact that it is on average more "risky" for students in neighborhood j to apply to an elite college, hence only the best students apply to H. This increases the admission threshold, which induces more self-selection in neighborhood i. At equilibrium, students in j end up more optimistic about their admission chances than students in i, yielding $q'_j/q'_i > q_j/q_i$ and $W'_j/W'_i > W_j/W_i$.

2.3.3 Policy Instruments

We discuss the effect of two possible policy interventions. The first one consists in imposing quotas, pre-defining the number of seats each neighborhood should have. The second one consists in changing the compositions of the two

neighborhoods by imposing some degree of mixing while leaving the equilibrium force determines the number of seats assigned to each neighborhood. When considering these interventions, we will discuss the effect in terms of welfare, in terms of expected quality of admitted students as well as a comparison of how the two neighborhoods benefit from the intervention.

Quotas. Here, we impose that the two neighborhoods should have a number of seats proportional to their size, i.e. $q_i = q_j$. We investigate the impact on welfare loss (compared to the first-best allocation). We show that quotas are a redistribution tool across neighborhoods, but do not always lead to welfare gains. Indeed, in the uniform case with small capacities at elite colleges quotas are welfare neutral.

PROPOSITION 8. *Consider two neighborhoods with costs uniformly distributed on* $[0, \overline{c}_i]$ *and* $[0, \overline{c}_j]$. *As* $q \longrightarrow 0$, *quotas have no effect on aggregate welfare at the first order.*

This neutrality result, however, seems specific to the uniform case. As the next proposition shows, if inequality across neighborhoods is initially very large then quotas can increase the quality of admitted students which can be increase aggregate welfare.

PROPOSITION 9. *Suppose that cost is zero for all students in neighborhood* i *and uniformly distributed on* $[0, \overline{c}_j]$ *in neighborhood* j. *As* $\overline{c}_j \longrightarrow 0$ *and* $q \longrightarrow 0$, *quotas increase aggregate welfare compared to the case without quotas*

The intuition is that, without quotas, many average-ability students from neighborhood i get admitted to an elite college because applications are costless from them, and many high-ability students from j self-select. With quotas, however, the best students from both groups get admitted which raises welfare. Moreover, as cost in the poor neighborhood vanishes, inefficient applications in this neighborhood have no impact on welfare.

In our simulations with uniform cost on $[0, 0.1]$ in i and $[0, 0.3]$ in j we do not see much effect on welfare and ability of quotas. However, quotas have significant a redistributive effect. They are useful to transfer welfare from neighborhood i to neighborhood j, as shown by the next figure. Without quotas,

the welfare in neighborhood j represents half of the welfare in neighborhood i for low capacity at elite colleges. Instead, with quotas, welfare in the two neighborhoods are roughly identical.

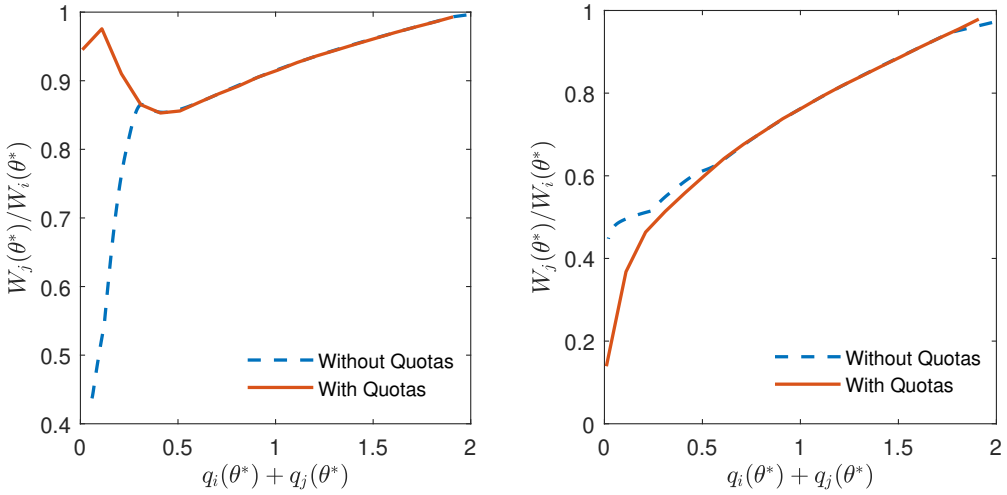


Figure 2.6: *Relative distribution of welfare across neighborhoods. (Left) Cost is uniformly distributed on* $[0, 0.1]$ *in* i *and* $[0, 0.3]$ *in* j*. (Right) Cost is uniformly distributed on* $[0, 0.5]$ *in* i *and* $[0.5, 1]$ *in* j*.*

As we mentioned in the previous section, when costs are high quotas may have counter-intuitive effects and reinforce inequalities. This is due to the fact that, in this case, the "disadvantaged neighborhood" already has more seats at elite colleges for small capacities q without quotas.

The idea that quotas induce a redistribution of welfare across neighborhoods is captured by the following result.

PROPOSITION 10. *Consider two neighborhoods with costs uniformly distributed on* $[0, \bar{c}_i]$ *and* $[0, \bar{c}_j]$*. Suppose that* $q < \max\{\sup_\theta \hat{q}_i(\theta), \sup_\theta \hat{q}_j(\theta)\}$*. With quotas, subjective admission chances decrease in the advantaged neighborhood, and increase in the disadvantaged neighborhood compared to the case without quotas.*

Mixed Neighborhoods. We investigate whether moving students from the poor neighborhood to the rich neighborhood (and vise versa) increases welfare. Unlike quotas which do not change students' social network, this intervention exactly aims at reducing inequalities of social capital. We consider random

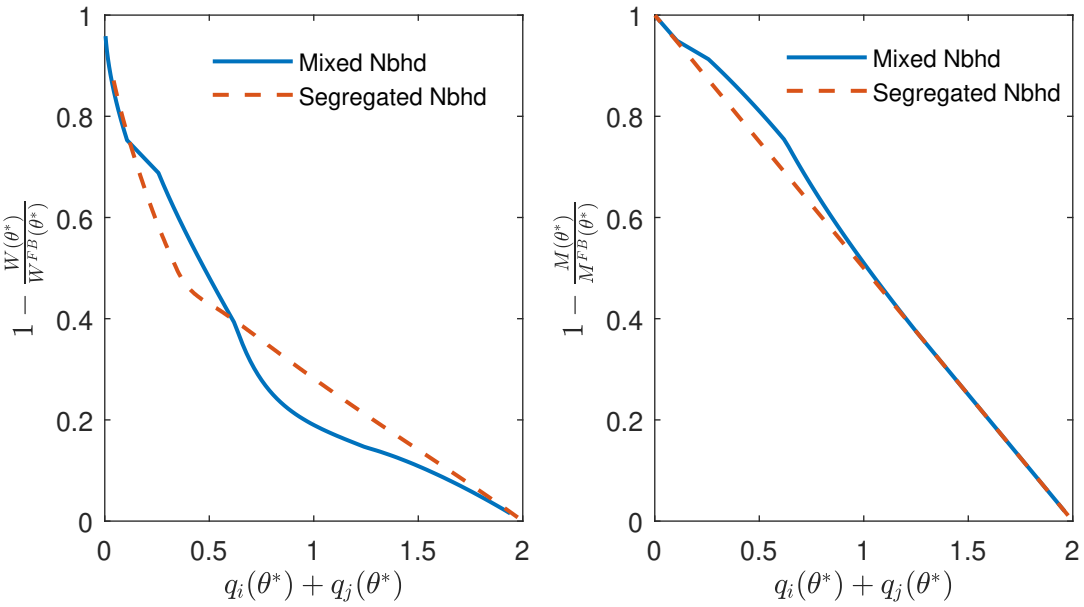


Figure 2.7: *Welfare loss and quality loss with respect to the first best allocation (i.e., rational expectations). Cost is uniformly distributed on* $[0, 0.5]$ *in* i *and* $[0.5, 1]$ *in* j.

reallocation, i.e. from two initial neighborhoods with cost distributions G_i and G_j we draw new neighborhoods from the following compound distributions:

$$\tilde{G}_i = \alpha G_i + (1 - \alpha) G_j$$
$$\tilde{G}_j = \alpha G_j + (1 - \alpha) G_i$$

The parameter α scales the equalization across neighborhoods: for $\alpha = 1$ there is no reallocation of students, and for $\alpha = \frac{1}{2}$ the new neighborhoods have equal cost distributions.

Our main result is that mixing is welfare neutral in the uniform case with small costs.

PROPOSITION 11. *Consider two neighborhoods with costs uniformly distributed on* $[0, \bar{c}_i]$ *and* $[0, \bar{c}_j]$. *Suppose that* $q < \max\{\sup_\theta \hat{q}_i(\theta), \sup_\theta \hat{q}_j(\theta)\}$. *Subjective admission chances* p_i, p_j *are independent of the degree of mixing* α. *Therefore, mixing is neutral on welfare and average quality.*

Affirmative Action. We now suppose that colleges provide a boost to the score of disadvantaged students. This is a fairly common policy instrument when schools rank students using a score that aggregates several characteristics. We assume that there is a cutoff c^* such that, if a student with cost higher than c^* applies to H, then the admission cutoff is $\theta^* - \kappa$ where $\kappa > 0$ is a constant chosen by the school. Instead, if a student with cost lower than c^* applies to H, then the admission cutoff is simply θ^*.

We show that affirmative action can decrease welfare when costs are small, as this policy leads to fewer admission of high-achieving advantaged students which is not compensated by greater admission of high-achieving disadvantaged students. This negative result, however, rely on small costs and may not hold for larger costs. Moreover, affirmative action still has a redistributive effect towards disadvantaged students.

PROPOSITION 12. *Suppose that cost is zero for all students in neighborhood i and uniformly distributed on $[0, \overline{c}_j]$ in neighborhood j. As $\overline{c}_j \longrightarrow 0$, affirmative action decreases aggregate welfare.*

2.3.4 Equilibrium Multiplicity

We conclude by investigating the role of equilibrium multiplicity on belief traps. In most of the analysis so far, we have assumed that cost and ability are independent. It turns out that this independence rules out equilibrium multiplicity.

PROPOSITION 13. *If cost and ability are independent $f(\theta, c) = h(\theta)g(c)$, then the local sampling equilibrium is unique.*

In reality, however, it could be that cost and ability are in fact correlated. When multiple neighborhoods compete for the same seats at elite colleges, this can lead to equilibria with belief traps in which one neighborhood takes all seats—even if all neighborhoods are ex-ante identical. For instance, this happens when ability and cost are positively correlated: if all seats are taken by students from neighborhood i, and low-cost low-ability students from neighborhood j are getting rejected, this is convincing evidence for high-ability students from neighborhood j that they admission chances are low.

PROPOSITION 14. *Suppose that in both neighborhoods there is a mass α of students with $(\theta, c) = (0, 0)$ and a mass $1 - \alpha$ of students with $(\theta, c) = (\varepsilon, \varepsilon)$ for $\varepsilon > 0$ small. There is an equilibrium in which all seats at H are taken by students from neighborhood i.*

2.4 Conclusion

We introduce a model of expectation formation in a career choice problem. Unlike the rational expectations framework, students have no prior information and no prior belief as to how elite colleges admit students. We assume instead that students non-parametrically estimate the distribution of outcomes conditional on actions by averaging past experiences from their peers with similar characteristics. Formally, we introduce a new solution concept—the local sampling equilibrium—in which players best respond to their subjective expectations, and expectations are consistent with the average outcomes of their peers. This provides a coherent framework for thinking the strategic interactions between expectation formation and the social environment.

We derive three main results. First, expectation formation leads to belief traps whereby high-achieving disadvantaged students self-select out of elite colleges, and average-ability advantaged students take their seats at elite colleges. This is due to the fact that average students create a strategic externality on high-achieving students by distorting their perceived admission chances toward the mean. This leads to multiple inefficiencies: on the supply side, high-achieving disadvantaged students go on the labor market instead of attending elite colleges, whereas low-achieving advantaged students spend resources applying to elite colleges even though their actual admission chances are zero. On the demand side, the pool of admitted students is of lower quality compared to the rational expectations benchmark.

Second, we show that a decrease in the average cost in one neighborhood has a negative impact on self-selection in *other* neighborhoods. This type of cross-neighborhood externality arises because rationing at elite colleges acts as a propagation mechanism of local demand shocks. Indeed, a reduction of cost in one neighborhood induces a higher admission cutoff, leading to a lower ad-

mission rates in other neighborhoods hence more self-selection. This suggests that growth inequality across locations disproportionately benefits advantaged neighborhoods at the expense of poor neighborhoods.

Finally, we show that quotas can mitigate the effects of neighborhood inequalities. Quotas reduce belief distortion leading to a better pool of applicants in the disadvantaged neighborhood. We argue that this might explain why there is little empirical support for the "mismatch hypothesis" which asserts that affirmative action policies results in minority students being admitted to colleges for which they are otherwise unqualified—leading to lower graduation rates and eventually harming minority students.

Application Cost

An alternative to the opportunity cost is to consider an application cost: all else equal, it is harder for disadvantaged students to apply to elite colleges because they don't have access to peers or professional who can help them in the process.

The payoffs are as follows:

- If student (θ, c) goes on the labor market L her utility is 0.
- If student (θ, c) applies to H and obtain a seat, her utility is $\theta - c$.
- If student (θ, c) applies to H but does not get a seat, she goes on the labor market and her utility is $-c$.

Student (θ, c) applies to H whenever $p(\theta)\theta - c \geq 0$, that is, whenever $c^H(\theta, p) \leq p(\theta)\theta$. Define welfare as

$$W(\sigma) = \int_{\theta^*}^{\overline{\theta}} \int_0^{c^H} \theta f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta - \int_0^1 \int_0^{c^H} c f(\theta, c) \, \mathrm{d}c \, \mathrm{d}\theta$$

It is readily verified that with one neighborhood: (i) the rational expectation equilibrium with application cost is identical than with opportunity cost, and (ii) the local-sampling equilibrium with application cost is identical than with

opportunity cost (up to the thresholds c^H). Therefore, with one neighborhood the analysis and the qualitative predictions are very similar.

With multiple neighborhoods, the welfare effect of policy instruments is similar with application cost and opportunity cost. For instance, quotas are welfare neutral with uniform cost.

PROPOSITION 15. *Consider two neighborhoods with costs uniformly distributed on* $[0, \overline{c}_i]$ *and* $[0, \overline{c}_j]$*. As* $q \longrightarrow 0$*, quotas have no effect on aggregate welfare at the first order.*

Proof of Proposition 15. First, we derive subjective admission chances in the case with quotas. We consider the following neighborhood specific quotas: $q_i = q_j = \frac{q}{2}$. The subjective admission chances for each neighborhoods write:

$$p_i = \sqrt{\overline{c}_i q} \qquad p_j = \sqrt{\overline{c}_j q}$$

The neighborhood specific admission cutoff $\tilde{\theta}_i^*$ solves

$$\int_{\tilde{\theta}_i^*}^{1} \frac{p_i}{\overline{c}_i} \theta \, \mathrm{d}\theta = \frac{q}{2} \iff \tilde{\theta}_i^* = \sqrt{1 - q\frac{\overline{c}_i}{p_i}}. \tag{8}$$

The admission cutoff in neighborhood j is similar, replacing p_i with p_j.

Second, we approximate $W(\theta^*)$ at the first order and show that it is independent of $\overline{c}_i$ and $\overline{c}_j$. As $q \longrightarrow 0$, we have

$$W_i(\theta^*) = p_i \frac{1 - (\theta^*)^3}{3\overline{c}_i} - \frac{p_i^2}{6\overline{c}_i}.$$

For $q \longrightarrow 0$, we make the following approximation: $(\theta^*)^3 \approx 1 - 3\sqrt{q\overline{c}_i}$. Therefore we obtain $W_i \approx \frac{5}{6}q$ at the first order. Hence, $W(\theta^*) \approx \frac{5}{3}q$ is independent of $\overline{c}_i, \overline{c}_j$. □

Proofs

Existence of Local Sampling Equilibria. Consider the following scheme:

$$p \longmapsto \sigma^{BR}(p, \cdot) \longmapsto b(\sigma^{BR}, \cdot) \longmapsto p(b)$$

By Tychonoff's theorem, the scheme is compact-valued $p(b) \in [0,1]^{\Theta}$. Hence to obtain a fixed point, we just need to prove that the scheme is continuous. Fix a subjective belief map $p : \Theta \longrightarrow [0,1]$. The action space is binary and the subjective admission chances p enter payoffs linearly, hence σ^{BR} is the following measurable threshold strategy:

$$\sigma^{BR}(p,\cdot) = \begin{cases} 1 & \text{if} \quad p(\cdot) \geq \gamma(\cdot) \\ 0 & \text{if} \quad p(\cdot) < \gamma(\cdot) \end{cases}$$

where $\gamma(\theta, c) = \frac{c}{\theta + c}$. Take any converging sequence $p_n \longrightarrow p$. We need to show that $p \longmapsto \sigma^{BR}(p,\cdot)$ is continuous in the L^1-weak topology, namely

$$\int \sigma^{BR}(p_n,(\theta,c))\,\mathrm{d}F \longrightarrow \int \sigma^{BR}(p,(\theta,c))\,\mathrm{d}F.$$

We have

$$\int \sigma^{BR}(p_n,(\theta,c))\,\mathrm{d}F = \int \mathbf{1}\left\{p_n(\theta) \geq \gamma(\theta,c)\right\}\,\mathrm{d}F.$$

Therefore, continuity follows from Lebesgue's dominated convergence theorem. We now show the continuity of $\sigma^{BR} \longmapsto b(\sigma^{BR},\cdot)$. By Berge's maximum theorem, $\sigma^{BR} \longmapsto b(\sigma^{BR},\cdot)$ is upper-hemicontinuous. The loss function $|\theta - \tilde{\theta}|$ is strictly quasi-convex, hence $\sigma^{BR} \longmapsto b(\sigma^{BR},\cdot)$ is continuous. Finally, the continuity of $b \longmapsto p(b)$ follows directly from the integrability of p together with the continuity of the functions $\max\{\cdot,\cdot\}$ and $\min\{\cdot,\cdot\}$. Therefore, by the Schauder fixed point theorem the set of local sampling equilibria is nonempty. □

Proof of Proposition 1. Fix the admission cutoff at $\theta^* = H^{-1}(1-q)$. If $\sigma^R(\theta,c) = 1$ for all $\theta > H^{-1}(1-q)$ and 0 otherwise, then the beliefs

$$p^R(\theta) = \begin{cases} 1 & \text{when} \quad \theta > H^{-1}(1-q) \\ 0 & \text{when} \quad \theta < H^{-1}(1-q) \end{cases}$$

are rationally consistent with σ^R by definition of θ^*. Given these subjective beliefs, $\sigma^R(\theta,c) = 1$ for all $\theta > H^{-1}(1-q)$ and 0 otherwise is optimal. Therefore,

(σ^R, p^R) is a rational expectations equilibrium.

We now prove uniqueness. Suppose that $\sigma(\theta, c) < 1$ for some (positive mass of) $\theta > \theta^*$ and $\sigma(\theta, c) > 0$ for some (positive mass of) $\theta < \theta^*$. By belief consistency, students with ability $\theta > \theta^*$ know that $p^R(\theta) = 1$ (i.e. they can obtain a seat at H for sure) hence they have a profitable deviation. □

Proof of Proposition 2. First we show that all students (θ, c) with ability $\theta > \theta^* = H^{-1}(1-q)$ and cost $c > c^H(\theta^*, p(\theta^*))$ self-select out of elite colleges. Student (θ, c) applies to H only if $p(\theta^*) \geq \frac{c}{\theta^*+c}$. As long as $q < 1$ and $\tau > 0$, we must have $p(\theta^*) < 1$ because the last admitted student (θ^*, c) includes rejected students in her sample. Therefore, as $\lim_{c\to\infty} \frac{c}{\theta^*+c} = 1$ there must exist a positive g-measure of costs such that $p(\theta^*) < \frac{c}{\theta^*+c}$ because g has full support on $\mathbb{R}_+$. This proves that self-selection arises in equilibrium.

Second, we show that students with ability $\theta < \tilde{\theta}^*$ and cost $c < c^H(\tilde{\theta}^*, p(\tilde{\theta}^*))$ apply to H but are rejected. Student (θ, c) with $\theta = \tilde{\theta}^* - \varepsilon$ for $\varepsilon > 0$ arbitrarily small applies to H only if $p(\theta^*) \geq \frac{c}{\tilde{\theta}^*+c}$. As long as $q < 1$ and $\tau > 0$, $p(\theta^*) > 0$ because this student includes in her sample admitted peers for ε small enough. Therefore, as $\lim_{c\to 0} \frac{c}{\tilde{\theta}^*+c} = 0$ there must exist a positive g-measure of costs such that $p(\tilde{\theta}^*) > \frac{c}{\tilde{\theta}^*+c}$ because g has full support on $\mathbb{R}_+$. This proves that inefficient applications arise in equilibrium. □

Proof of Proposition 3. We can rewrite the implicit equation for subjective beliefs as follows:

$$p - \frac{1}{\tau}\int_{\underline{\theta}}^{\overline{\theta}} \mathbf{1}_{\{\theta - b(\theta,\sigma) < \tilde{\theta} < \theta + b(\theta,\sigma)\} \cap \{\tilde{\theta} > \tilde{\theta}^*\}} G\left(\frac{p(\tilde{\theta})}{1 - p(\tilde{\theta})}\right) \mathrm{d}H(\tilde{\theta}) = 0 \tag{9}$$

We first consider the case in which $\tau \to 1$. By definition of $b(\theta, \sigma)$ we have $\lim_{\tau\to 1}\{\theta - b(\theta,\sigma) < \tilde{\theta} < \theta + b(\theta,\sigma)\} \supseteq \{\tilde{\theta} > \tilde{\theta}^*\}$. Therefore,

$$\lim_{\tau\to 1}\left[p(\theta) - \frac{1}{\tau}\int_{\underline{\theta}}^{\overline{\theta}} \mathbf{1}_{\{\theta - b(\theta,\sigma) < \tilde{\theta} < \theta + b(\theta,\sigma)\}} G\left(\frac{p(\tilde{\theta})}{1 - p(\tilde{\theta})}\right) \mathrm{d}H(\tilde{\theta})\right] = 0$$

$$\iff p = \int_{\underline{\theta}}^{\overline{\theta}} G\left(\frac{p}{1-p}\right) \mathrm{d}H(\tilde{\theta})$$

where the second line uses the fact that, as $\tau \to 1$, the subjective probability becomes independent of θ.

We now consider the case $\tau \to 0$. There are two cases to consider.

Case 1: There exists τ_* small enough such that $\theta + b(\theta, \sigma) < \tilde{\theta}^*$. Then we have $\lim_{\tau \to 1}\{\theta - b(\theta, \sigma) < \tilde{\theta} < \theta + b(\theta, \sigma)\} \cap \{\tilde{\theta} > \tilde{\theta}^*\} = \emptyset$. Hence taking the integral in equation (9) is zero, and we directly have that $p(\theta) = 0$.

Case 2: There exists τ^* small enough such that $\theta - b(\theta, \sigma) > \tilde{\theta}^*$. Then we have $\lim_{\tau \to 1}\{\theta - b(\theta, \sigma) < \tilde{\theta} < \theta + b(\theta, \sigma)\} \subseteq \{\tilde{\theta} > \tilde{\theta}^*\}$. Therefore,

$$\lim_{\tau \to 0} \left[p(\theta) - \frac{1}{\tau} \int_{\underline{\theta}}^{\overline{\theta}} \mathbf{1}_{\{\theta - b(\theta,\sigma) < \tilde{\theta} < \theta + b(\theta,\sigma)\}} G\left(\frac{p(\tilde{\theta})}{1 - p(\tilde{\theta})} \right) \mathrm{d}H(\tilde{\theta}) \right] = 0$$

Take $p(\theta) = 1$ and using the fact that $\lim_{x \to \infty} G(x) = 1$ we can rewrite the above equation as follows:

$$\lim_{\tau \to 0} \left[1 - \frac{1}{\tau} \int_{\theta - b(\theta,\sigma)}^{\theta + b(\theta,\sigma)} h(\tilde{\theta}) \, \mathrm{d}\tilde{\theta} \right] = 0$$

By L'Hospital's rule and Leibniz integral rule,

$$\lim_{\tau \to 0} \frac{\int_{\theta - b(\theta,\sigma)}^{\theta + b(\theta,\sigma)} h(\tilde{\theta}) \, \mathrm{d}\tilde{\theta}}{\tau} = \lim_{\tau \to 0} \Big[h(\theta + b(\theta, \sigma)) + h(\theta - b(\theta, \sigma)) \Big] \frac{\partial b(\theta, \sigma)}{\partial \tau} \tag{10}$$

By definition, $b(\theta, \sigma)$ is the smallest $b > 0$ that solves:

$$\int_{\theta - b}^{\theta + b} f(\theta) \, \mathrm{d}\theta > \tau \iff \underbrace{H(\theta + b) - H(\theta - b) - \tau}_{= \Phi(b, \tau)} > 0$$

We apply the implicit function theorem to obtain the derivative of $b(\theta, \tau)$:

$$\frac{\partial \Phi}{\partial b} \frac{\partial b}{\partial \tau} + \frac{\partial \Phi}{\partial \tau} = 0 \iff \frac{\partial b}{\partial \tau} = \frac{1}{h(\theta + b) + h(\theta - b)}$$

Substituting this expression in equation (10) concludes the proof. □

Proof of Proposition 6. When $\lim_n \tau_i^n = 0$, we already showed that $p_i = \mathbf{1}\{\theta_i \geq \theta^*\}$, hence $\{i : \sigma_i = 1\} = \{i : \theta_i \geq \theta^*\}$ and $q_i = |\{i : \sigma_i = 1 \text{ and } \theta_i \geq \theta^*\}| = |\{i :$

$\theta_i \geq \theta^*\}|$. By contradiction, suppose that $q_i < q_j$. Then $|\{j : \sigma_j = 1 \text{ and } \theta_j \geq \theta^*\}| > |\{i : \theta_i \geq \theta^*\}|$. Note that $\{j : \sigma_j = 1 \text{ and } \theta_j \geq \theta^*\} \subseteq \{j : \theta_j \geq \theta^*\}$, hence $|\{j : \theta_j \geq \theta^*\}| \geq |\{j : \sigma_j = 1 \text{ and } \theta_j \geq \theta^*\}|$. But then, $|\{j : \theta_j \geq \theta^*\}| > |\{i : \theta_i \geq \theta^*\}|$, which contradicts the fact that $f_i = f_j$. □

Proof of Proposition 7. When $G_i = \delta_0$ we have that $\sigma_i = 1$ for all i. Hence, $q_i = |\{i : \sigma_i = 1 \text{ and } \theta_i \geq \theta^*\}| = |\{i : \theta_i \geq \theta^*\}|$. We conclude using the same reasoning as in the proof of Proposition 6. □

Proof of Proposition 8. First, we derive subjective admission chances in the case with quotas. We consider the following neighborhood specific quotas: $q_i = q_j = \frac{q}{2}$. The subjective admission chances (computed in Section 3.1) for each neighborhoods write:

$$p_i = -\frac{\overline{c}_i q}{2} + \sqrt{\frac{\overline{c}_i^2 q^2}{4} + \overline{c}_i q} \qquad p_j = -\frac{\overline{c}_j q}{2} + \sqrt{\frac{\overline{c}_j^2 q^2}{4} + \overline{c}_j q}$$

The neighborhood specific admission cutoff $\tilde{\theta}_i^*$ solves

$$\int_{\tilde{\theta}_i^*}^1 \frac{p_i}{\overline{c}_i(1-p_i)} \theta \, \mathrm{d}\theta = \frac{q}{2} \iff \tilde{\theta}_i^* = \sqrt{1 - q\overline{c}_i \frac{1-p_i}{p_i}}. \tag{11}$$

The admission cutoff in neighborhood j is similar, replacing p_i with p_j.

Second, we approximate $W(\theta^*)$ at the first order and show that it is independent of $\overline{c}_i$ and $\overline{c}_j$. As $q \longrightarrow 0$, we have

$$W_i(\theta^*) = \frac{p_i}{1-p_i} \frac{1-(\theta^*)^3}{3\overline{c}_i} - \left(\frac{p}{1-p}\right)^2 \frac{(\theta^*)^3}{6\overline{c}_i}.$$

Again for $q \longrightarrow 0$, we make the following approximations: $p_i \approx \sqrt{2q\overline{c}_i}$, $\frac{p_i}{1-p_i} \approx p_i$ and $(\theta^*)^3 \approx 1 - \frac{3}{2}\sqrt{2q\overline{c}_i}$. Therefore we obtain $W_i \approx \frac{2}{3}q$ at the first order. Hence, $W(\theta^*) \approx \frac{4}{3}q$ is independent of $\overline{c}_i, \overline{c}_j$. □

Proof of Proposition 9 All students in neighborhood i are indifferent hence apply to H. Then without quotas all students $\{(\theta, c) : \theta \geq \theta^*\}$ are admitted to H in neighborhood i. In neighborhood j, all students $\left\{(\theta, c) : \theta \geq \theta^* \text{ and } c \leq \frac{p_j}{\overline{c}_j(1-p_j)}\theta\right\}$

are admitted to H. Overall welfare is

$$W(\theta^*) = \underbrace{\int_{\theta^*}^{1} \theta \, \mathrm{d}\theta}_{i} + \underbrace{\int_{\theta^*}^{1} \int_{0}^{c^H} \theta \, \mathrm{d}c \, \mathrm{d}\theta - \int_{0}^{\theta^*} \int_{0}^{c^H} c g_j(c) \, \mathrm{d}c \, \mathrm{d}\theta}_{j} \tag{12}$$

With quotas, each neighborhood has $q/2$ reserved seats. In neighborhood j, quotas must increase subjective admission chances p_j. Indeed, as $\overline{c}_j \longrightarrow 0$ the best students in both neighborhoods apply to H hence the admission cutoff solves $2(1 - \theta^*_{quotas}) = q$. Instead, without quotas the admission cutoff solves $1 - \theta^* = q$, which is strictly smaller than with quotas. This raises the quality of admitted students in both neighborhoods, which increases the first two terms in the welfare equation (12). Now the third term vanishes as $\overline{c}_j \longrightarrow 0$, which yields the result. □

Proof of Proposition 10. We already derived subjective admission chances in the case of quotas in Proposition 8. Therefore, we derive them in the case without quotas.

When there are no quotas, both neighborhoods compete for the same q seats. Subjective admission chances in neighborhoods i are obtained by dividing the number of seats by the mass of applicants in this neighborhood:

$$p_i = \frac{q_i}{\int_0^1 \min\left\{\frac{p_i}{\overline{c}_i(1-p_i)}\theta, 1\right\} \mathrm{d}\theta} \tag{13}$$

where $q_i + q_j = q$ are the seats taken by students from neighborhoods i and j in equilibrium. We consider first the case in which the minimum does not bind in both neighborhoods. The above equation rewrite:

$$p_i = 2q_i\overline{c}_i\frac{1-p_i}{p_i} \tag{14}$$

The market clearing condition in neighborhood i writes:

$$\int_{\tilde{\theta}_i^*}^{1} \min\left\{\frac{p_i}{\overline{c}_i(1-p_i)}\theta, 1\right\} \mathrm{d}\theta = q_i \iff 1 - \tilde{\theta}_i^{*2} = 2q_i\overline{c}_i\frac{1-p_i}{p_i}$$

Together with the fact that admission cutoffs must be equal across neighborhoods $\tilde{\theta}_i^* = \tilde{\theta}_j^*$, this shows that subjective beliefs are identical $p_i = p_j = p$.

Using the market clearing condition together with the identity $q = q_i + q_j$ we obtain the number of seats taken by each neighorhoods in equilibrium:

$$q_i = q\frac{\overline{c}_j}{\overline{c}_i + \overline{c}_j}$$

Solving the quadratic form (14) and substituting the expression for q_j yields a closed form solution for subjective beliefs:

$$p^{\textit{no quotas}} = -q\frac{\overline{c}_i\overline{c}_j}{\overline{c}_i + \overline{c}_j} + \sqrt{\left(q\frac{\overline{c}_i\overline{c}_j}{\overline{c}_i + \overline{c}_j}\right)^2 + 2q\frac{\overline{c}_i\overline{c}_j}{\overline{c}_i + \overline{c}_j}}.$$

It can be verified that $p_i > p^{\textit{no quotas}} > p_j$. □

Proof of Proposition 11. Subjective admission chances in neighborhoods i are obtained by dividing the number of seats by the mass of applicants in this neighborhood:

$$p_i = \frac{q_i}{\int_0^1 \alpha \min\left\{\frac{p_i}{\overline{c}_i(1-p_i)}\theta, 1\right\} + (1-\alpha)\min\left\{\frac{p_i}{\overline{c}_j(1-p_i)}\theta, 1\right\} \mathrm{d}\theta} \tag{15}$$

where $q_i + q_j = q$ are the seats taken by students from neighborhoods i and j in equilibrium. We consider first the case in which the minimum does not bind in both neighborhoods. The above equation rewrite:

$$p_i = 2q_i\frac{1-p_i}{p_i}\left[\frac{\alpha}{\overline{c}_i} + \frac{1-\alpha}{\overline{c}_j}\right]^{-1} \tag{16}$$

and in neighborhood j:

$$p_j = 2q_j\frac{1-p_j}{p_j}\left[\frac{\alpha}{\overline{c}_j} + \frac{1-\alpha}{\overline{c}_i}\right]^{-1} \tag{17}$$

The market clearing condition in neighborhood i writes:

$$\int_{\tilde{\theta}_i^*}^1 \alpha \min\left\{\frac{p_i}{\bar{c}_i(1-p_i)}\theta, 1\right\} + (1-\alpha)\min\left\{\frac{p_i}{\bar{c}_j(1-p_i)}\theta, 1\right\} \mathrm{d}\theta = q_i$$
$$\iff 1 - \tilde{\theta}_i^{*2} = 2q_i \frac{1-p_i}{p_i}\left[\frac{\alpha}{\bar{c}_i} + \frac{1-\alpha}{\bar{c}_j}\right]^{-1}$$

Together with the fact that admission cutoffs must be equal across neighborhoods $\tilde{\theta}_i^* = \tilde{\theta}_j^*$, this shows that subjective beliefs are identical $p_i = p_j = p$. Solving for q_i yields

$$q_i = q\left(\frac{\alpha}{\bar{c}_j} + \frac{1-\alpha}{\bar{c}_i}\right)^{-1}\left[\left(\frac{\alpha}{\bar{c}_j} + \frac{1-\alpha}{\bar{c}_i}\right)^{-1} + \left(\frac{\alpha}{\bar{c}_i} + \frac{1-\alpha}{\bar{c}_j}\right)^{-1}\right]^{-1}$$

Solving the quadratic form (17) yields a closed form solution for subjective beliefs:

$$p = \frac{-q_i + \sqrt{q_i^2 + 2q_i\left(\frac{\alpha}{\bar{c}_i} + \frac{1-\alpha}{\bar{c}_j}\right)}}{\left(\frac{\alpha}{\bar{c}_i} + \frac{1-\alpha}{\bar{c}_j}\right)}.$$

which is constant in α. Therefore, for small q welfare and average quality is independent of α. □

Proof of Proposition 12. All students in neighborhood i are indifference hence apply to H. Costs in neighborhood i are all below c^*, hence the regular cutoff applies and all students $\{(\theta, c) : \theta \geq \theta^*\}$ are admitted to H. In neighborhood j, all students $\{(\theta, c) : \theta \geq \theta^* \text{ and } c < c^*\}$ who apply to H are admitted, as well as all students $\{(\theta, c) : \theta \geq \theta^* - \kappa \text{ and } c^* \leq c \leq \frac{p_j}{\bar{c}_j(1-p_j)}\theta\}$ who apply to H. Overall welfare with affirmative action is

$$W(\theta^*) = \int_{\theta^*}^1 \theta \,\mathrm{d}\theta + \int_{\theta^*}^1 \int_0^{\min\{c^*, c^H\}} \theta \,\mathrm{d}c \,\mathrm{d}\theta + \int_{\theta^*-\kappa}^1 \int_{\min\{c^*, c^H\}}^{\min\{\bar{c}_j, c^H\}} \theta \,\mathrm{d}c \,\mathrm{d}\theta$$
$$- \int_0^{\theta^*} \int_0^{\min\{c^*, c^H\}} cg(c) \,\mathrm{d}c \,\mathrm{d}\theta - \int_0^{\theta^*-\kappa} \int_{\min\{c^*, c^H\}}^{\min\{\bar{c}_j, c^H\}} cg(c) \,\mathrm{d}c \,\mathrm{d}\theta \qquad (18)$$

Without affirmative action, welfare is

$$W(\theta^*) = \int_{\theta^*}^{1} \theta \, \mathrm{d}\theta + \int_{\theta^*}^{1} \int_{0}^{\min\{\bar{c}_j, c^H\}} \theta \, \mathrm{d}c \, \mathrm{d}\theta - \int_{0}^{\theta^*} \int_{0}^{\min\{\bar{c}_j, c^H\}} cg(c) \, \mathrm{d}c \, \mathrm{d}\theta \qquad (19)$$

Note that affirmative action has no effect on subjective admission chances in neighborhood i fixing the cutoff θ^*. Instead, in neighborhood j this makes high achieving students more optimistic about their admission chances. It must be that θ^* increases for the market clearing condition $q_1 + q_2 = q$ to be satisfied. Hence, the first two terms in equation (18) decrease. Now, as $\bar{c}_j \longrightarrow 0$, the last three terms in equation (18) vanish. Therefore, overall welfare decreases. □

Proof of Proposition 4. By contradiction, suppose that there exist two equilibria A and B. Without loss of generality suppose that we have $p^A > p^B$. By independence, the set of students who apply to H in equilibrium A is a superset of the set of students who apply to H in equilibrium B because subjective admission chances are higher in A. Note however that p is the ratio of seats to the number of applicants, i.e. $p = q \Big/ \left[\int_0^1 \min \left\{ \frac{\frac{p}{1-p}\theta}{\bar{c}}, 1 \right\} \mathrm{d}\theta \right]$. Hence, we must have $p^A < p^B$, a contradiction. □

Proof of Proposition 14. Suppose that all seats are taken by students from neighborhood i ($q_i = q$), and suppose that ε is small enough such that all seats are taken by high-ability students, namely $p_i\varepsilon - (1 - p_i)\varepsilon > 0$ holds. Therefore, the admission cutoff satisfies $\theta^* > 0$. Suppose that all low-ability students from neighborhood j apply to H but are being rejected because $\theta_j = 0 < \theta^*$. Then we have $p_j = 0$, and no high-ability student in neighborhood j applies to H. There are no profitable deviations and beliefs are consistent, hence this is a local sampling equilibrium. □

Chapter 3

Revealed Deliberate Preference Change[1]

Understanding how individuals change their behavior is critical for social sciences. Economists traditionally argue that decision makers (DMs) are Bayesian; that is, they adapt their behavior by updating their beliefs about the economic environment. Although this mechanism has proved powerful and normatively appealing, a wide range of phenomena seem better described with preference change because they involve values such as fairness, conservatism, etc. For instance, Barrera et al. (2020) show experimentally that exposure to fake news about the European refugee crisis increases voting intentions toward far-right politicians. Importantly, fact checking modifies voters' beliefs but not voting intentions. One explanation they put forward is the saliency towards the issue raised by the politician, that may alter voters' awareness, and as a consequence their preferences. Another example is the expansion of abortion rights in western societies—along with its economical and political implications—that is more plausibly due to the diffusion of new values such as women rights than to changing beliefs on some underlying state of the world.

[1]This chapter is joint with Niels Boissonnet and Alexis Ghersengorin. We are grateful to Douglas Bernheim, Yves Breitmoser, Simone Cerreia-Vioglio, Franz Dietrich, Marco Mariotti, Pietro Ortoleva, Jean-Marc Tallon, Al Roth, Ariel Rubinstein and seminar participants at Bielefeld, CREST, PSE, Stanford and TUS-VI for helpful conversations and comments. A. Ghersengorin thanks ANR-17-CE26-0003 for its support. S. Gleyze acknowledges the support of the EUR grant ANR-17-EURE-0001.

Modeling preference changes raises two challenges: first, the lack of normative foundations compared to Bayesian updating; second the lack of testability of the model. To fill these gaps, we propose and axiomatize two testable normative principles: a *principle of sufficient reason* and a *principle of deliberation*. To express these normative principles, we use the attribute-based approach which adds structure to the set of alternatives. Our primitive is the observation of successive preferences on the alternatives, as well as the attributes of each alternative. This allows us to reveal DM's reasoning behind preference changes and is sufficiently tractable for applied and empirical work. In doing so, we make progress toward a testable and normatively founded model of preference change.

The principle of sufficient reason states that DM changes her preferences if and only if it can be justified by an attribute of the alternative that is made relevant or irrelevant. For instance, if an employer notices that her hiring decision is based on the attribute "gender", she might make this attribute irrelevant in the future to stop being discriminatory.[2] Formally, this translates into an identification axiom called Restricted Reversals, which guarantees that choice reversals must be induced by some attributes becoming (ir)relevant.

The principle of deliberation states that DM should not make mistakes (from her perspective) when changing preferences—that is, she cannot change her mind twice regarding an attribute if no additional event occurred meanwhile. Otherwise, this would indicate that she fails to deliberate and lacks internal consistency. Formally, this translates into an acyclicity axiom, which guarantees that if an attribute becomes relevant and then irrelevant it must be explained by *other* attributes becoming (ir)relevant meanwhile.

Our main representation theorem states that Restricted Reversals and Acyclicity hold if and only if (i) preferences are represented by the maximization of an ordering on the alternatives' attributes—we call it the attribute ordering—, and (ii) preference changes are explained by the maximization of an ordering on preferences themselves—we call it the meta-preference. Preference changes take the following form: whenever DM becomes aware of an attribute—through

[2] *Implicit* discrimination would also imply that the attribute "gender" is *relevant*. Therefore, an attribute can be relevant even if DM does not consciously use this attribute.

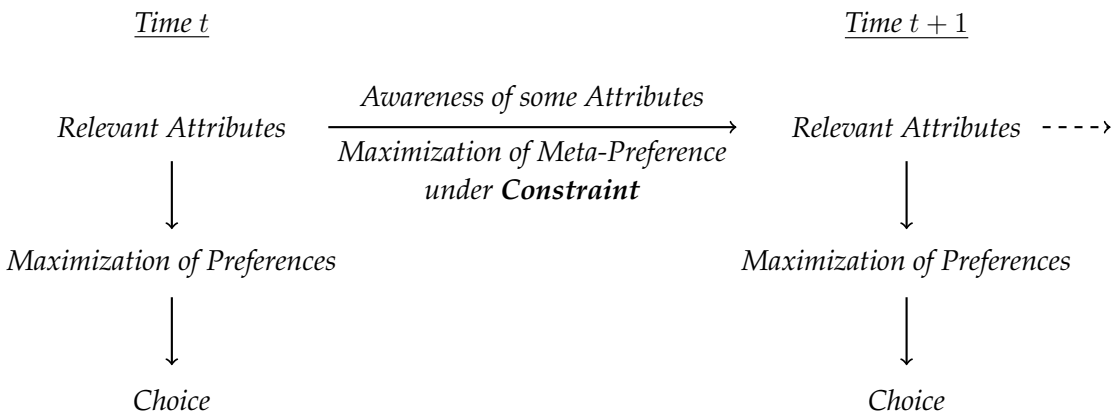


Figure 3.1: *The Dynamics of Deliberate Preference Change.*

education, social interactions, medias or introspection—she can decide to make it relevant or irrelevant for the next period, inducing a preference change. The succession of such changes is consistent with the maximization of a *meta-preference* relation, capturing DM's moral values, motivated reasoning, social objectives, norms, etc. Therefore, the reasoning behind preference changes is revealed through the meta-preference relation and the sequence of awareness. Such a sequence represents DM's constraint regarding which preferences are reachable at each period. The existence of such a constraint follows from the principle of deliberation and the observation of multiple choice reversals. Indeed, would DM be unconstrained in the maximization of her meta-preference she would directly reach her most preferred set of relevant attributes and never change preferences again. Note that the attribute ordering remains stable, only the set of relevant attributes changes; this implies that if DM deems relevant the same set of attributes from one period to another, she must make exactly the same choices.[3] See Figure 3.1 for a representation of the model.

We then provide a uniqueness result. Both the attribute ordering and the meta-preference are unique up to arbitrary completion. That is, if two distinct attribute orderings (resp. meta-preferences) rationalize the observed choices, they rank differently only irrelevant pairs of attribute combinations (resp. pref-

[3]We discuss why it would be problematic that DM changes her "taste" towards the attributes in Section 2.5.

erences). Furthermore, if two sequences of awareness represent DM's constraint on meta-choices, their intersection does too. We, however, stress that the sequence of relevant attributes is not uniquely identified in general. Hence, we investigate specific conditions that make this sequence set-identified or point-identified.

We then investigate a particular type of meta-preference (i.e., a particular type of reasoning) in which DM chooses the preferences that maximize her underlying utility. This captures *motivated preference change* in which DM's evaluation of the attributes is guided by her own-interest alone. We show that motivated preference change provides new insights on the formation of political preferences. For instance, if two voters with identical preferences become aware of the same attributes in a different order, they can end up endorsing antagonistic views. Whether a voter becomes aware that a politician is corrupted before or after learning his political affiliation can lead to very different outcomes: in the latter case, the voter might ignore this attribute because it undermines the view of her preferred candidate. This type of path-dependent motivated reasoning is specific to our model and provides empirically testable implications.

Our contribution is threefold: first, we show that models incorporating preference changes can have empirical content and normative foundations. Second, our model suggests that choice reversals need not be irrational, and may reflect DM aligning her choice behavior with her values. Though not all choice reversals are consistent with our model: any rational choice reversal must break (or create) indifference with respect to other pairs of alternatives that share the same attribute, which indicates that this attribute becomes relevant (resp. irrelevant). This is a necessary condition for preference change to be induced by a coherent reasoning from DM. Finally, we illustrate the explanatory power of our model through an application.

Related Literature. The idea of representing objects by their attributes goes back to Lancaster (1966). Moreover, we draw on an important literature on reason-based theories of choice, most notably Simonson (1989), Shafir et al. (1993), Tversky and Simonson (1993), and Dietrich and List (2013a, 2016). Bois-

sonnet (2019) provides a decision theoretic characterization of our model, and Dietrich and List (2013b) propose a related theory of non-informational preference change. Our paper should be seen as the first counterpart of these models within the revealed preference theory.

We also emphasize that there is an important literature on "changing tastes" understood as time inconsistency. Strotz (1955) is the first to uncover the problem of consistent planning and to investigate how should individuals with non-exponential discounting make dynamically consistent choices. Gul and Pesendorfer (2001, 2005) and Dekel et al. (2009) provide behavioral foundations of preferences for commitment, namely choosing a smaller choice set for one's future self to avoid temptation. The main differences with our paper is that they consider deviations between expected behavior and actual behavior which are typically *not deliberate* (inconsistent) from the point of view of past selves. Instead, we look at preference changes that are deliberate but completely myopic, meaning that DM is unaware that she may change preferences in the future. Moreover, they typically look at preferences over menus whereas we consider preferences on alternatives only. The closest paper in this literature to our own is Nehring (2006) who studies the revealed preference implications of second-order preferences as a self-control mechanism. The main differences with our paper are that he considers preferences over menus whereas we deal with preferences over alternatives, he does not introduce attributes, and the second order preferences act exclusively as a self-control mechanism whereas our meta-preference relation is completely general.

Our work relates on the literature on conflicting motivations—or justifiable choices—as we also obtain a representation with several (more precisely two) orderings. See among other contributions Kalai et al. (2002), Heller (2012), De Clippel and Eliaz (2012), Cherepanov et al. (2013), Dietrich and List (2016) and Ridout (2021). Despite this similarity, these works focus on static choice data that violates the usual rational requirements—namely the Weak Axiom of Revealed Preferences (WARP) or the Independence of Irrelevant Alternatives Axiom (IIA)—, whereas in our work, the choice data consists in an ordered sequence of choices on the same collection of menus of options. We explore two distinct situations, one in which within-period choices are represented by

not necessarily transitive binary relations, one in which within-period choices satisfy WARP. We focus on the irregularities in choices that arise between periods, hence the reversals can happen on the same menus. Furthermore, the time structure is used to rationalize the successive changes as being guided by a meta-maximization.

In the applied theory literature, the closest paper is Bernheim et al. (2021). Their model and ours share two important ideas. First, they argue that DM can choose "worldviews" which determine her valuation of future consumption streams. This is related to our concept of relevant attributes. Second, in their model DM is constrained by her "mindset flexibility" when changing worldviews. This echoes our constraint on awareness. For the purpose of falsification, our model makes some simplifications: in their model DM anticipates her preference change, and they allow for convex combinations of worldviews. Despite the differences in modelling assumptions, their paper is complementary with ours as we focus on the identification and falsification of deliberate preference changes. Other models of chosen preferences include Becker and Mulligan (1997), Akerlof and Kranton (2000), Palacios-Huerta and Santos (2004).

3.1 Deliberate Preference Change

3.1.1 Preliminaries

There is finite set X of **alternatives**, that are defined by their **attributes**. Formally, there are K attributes and an alternative is a vector $\mathbf{x} = (x^1, \ldots, x^K)$ in the vector space $\mathbb{R}^K$ whose k^{th}-coordinate describes the value x^k of the attribute k. For any subset $M \subseteq \{1, \ldots, K\}$, denote $\mathbf{x}^M = (x^k)_{k \in M}$ and $\mathbf{x}^{-M} = (x^k)_{k \notin M}$. If $x^k = 0$ for some k this is interpreted as $\mathbf{x}$ not possessing this attribute. The analyst observes (i) the value of each attribute for all alternatives, and (ii) choices over options for T periods of time. The latter are represented by a sequence of complete orders $(\succsim_t)_{t=1,\ldots,T}$, where $\succ_t$ and $\sim_t$ denote the asymmetric and symmetric parts, respectively. For the first part of the analysis, we do not require each $\succsim_t$ to be transitive. We investigate the implications of transitivity within periods —that is, DM's choices satisfy WARP— in section 3.1.6.

Example 1: Labor Market Discrimination. *An employer wants to hire a worker. Her decision is based on the resume of each candidate that provides information on three attributes: (1) "education", (2) "experience", and (3) "gender" (1 for female and 0 for male). Therefore, a female college-educated worker entering the labor market is represented by* $\mathbf{x} = (4, 0, 1)$*, while a male non-educated worker with ten years of experience is represented by* $\mathbf{y} = (0, 10, 0)$*.*

3.1.2 Revealed Relevant Attributes

The attribute-based approach allows us to identify which attributes drive DM's choice behavior. These "relevant attributes" are easy to identify when the choice set X is sufficiently rich: the attribute k is revealed relevant at t if there is a pair of alternatives $\mathbf{x}$ and $\mathbf{y}$ that only differ on the k^{th}-dimension ($\mathbf{x}^{-k} = \mathbf{y}^{-k}$) and such that $\mathbf{x} \not\sim_t \mathbf{y}$. In this case, we are sure that DM uses attribute k in her decision making. This richness assumption—that we can always find two alternatives that differ only on one dimension—would be too restrictive, however. For instance, it is violated in the Example 1. Therefore, we introduce a weaker notion of richness and show how to identify the revealed relevant attributes. We illustrate the construction using our running example, and then provide a formal definition.

Richness assumption. *For all* $\mathbf{x}, \mathbf{y} \in X$ *that differ only on a subset* M *of* n *attributes, there is a sequence of alternatives* $\mathbf{z}_1, \ldots, \mathbf{z}_n \in X$ *such that* $\mathbf{z}_1 = \mathbf{x}$*,* $\mathbf{z}_n = \mathbf{y}$ *and* $\mathbf{z}_i^{-k} = \mathbf{z}_{i+1}^{-k}$ *for some* $k \in M$*, for all* $i = 1, \ldots, n-1$*.*

This assumption states that for any pair of options $\mathbf{x}, \mathbf{y}$, we can find a chain of alternatives that differs only on one dimension and that connects $\mathbf{x}$ to $\mathbf{y}$.

Example 1 (continued): *Suppose there are only two candidates* $\mathbf{x} = (4, 0, 1)$*,* $\mathbf{z} = (4, 2, 0)$ *and* $\mathbf{z} \succ_t \mathbf{x}$*. The idea is to identify a set of attributes* $M \subset \{1, 2, 3\}$ *that has to be relevant to explain this strict preference. From* $\mathbf{z} \succ_t \mathbf{x}$*, we can conclude that* $M = \{2, 3\}$ *is revealed relevant because (i) the alternatives differ on* M *and are identical outside of* M*, and (ii) there is no pair of alternatives that differ on a strict subset of* M *and are ranked strictly. The second point captures conservatism in our*

definition of revealed relevant attributes: if we cannot disentangle which attributes drive DM's behavior exactly, we keep all attributes in M. The following definition formalizes points (i) and (ii).

DEFINITION 1— REVEALED RELEVANT ATTRIBUTE. *A set M of attributes is **revealed relevant** at period t if:*

(i) there exists $\mathbf{x}, \mathbf{y} \in X$ with $\mathbf{x}^{-M} = \mathbf{y}^{-M}$ and $x^k \neq y^k$ for every $k \in M$, such that $\mathbf{x} \not\sim_t \mathbf{y}$;

(ii) for every $M' \subsetneq M$ and every $\mathbf{w}, \mathbf{z} \in X$ with $\mathbf{w}^{-M'} = \mathbf{z}^{-M'}$, $\mathbf{w} \sim_t \mathbf{z}$.

Let P_t denote the collection of sets of revealed relevant attributes at period t. We denote $\mathbf{m}_t \in \{0,1\}^K$ the **vector of revealed relevant attributes** such that $m_t^k = 1$ if $k \in \bigcup_{M \in P_t} M$ and $m_t^k = 0$ otherwise.[4] Note that when $\succsim_t$ is transitive, we can restrict attention to singletons $M = \{k\}$ in the previous definition —see Lemma 1 in the Appendix. Moreover, our construction of revealed relevant attributes is valid even when the set of alternatives is not rich—but without richness, the model may not be exactly identified as we discuss in Section 3.1.7.

We want to emphasize that an attribute can be revealed relevant, yet DM might be unaware that it causes her behavior. For instance, it is well known that implicit discrimination can have a strong impact on job performance (Bertrand et al., 2005; Glover et al., 2017; Bertrand and Duflo, 2017).

3.1.3 Principle of Sufficient Reason

We impose the following principle of sufficient reason: DM changes preferences if and only if the revealed relevant attributes change. The interpretation is that DM does not "wake up" with different preferences but must be able to justify her new preferences by making some attributes relevant or irrelevant. We view this as a normative principle: unjustified changes would not be normatively compelling.

[4] Our definition of revealed relevant attributes is analogous to the definition of a non-null state in expected utility theory (taking the attributes as states and the alternatives as acts).

Formally, the axiom states that if two alternatives $\mathbf{x}$ and $\mathbf{x}'$ have the same relevant attributes between periods t and t'—namely, if $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ where $\circ$ denotes the element-wise (Hadamard) product—DM should rank consistenly $\mathbf{x}$ against the other alternatives in period t and $\mathbf{x}'$ against the other alternatives in period t'.

RESTRICTED REVERSALS. *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals if for any t, t', and for any $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'}$,*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_{t'} \mathbf{y}'.$$

Example 1 (continued). *Consider three candidates $\mathbf{x} = (6, 2, 1), \mathbf{x}' = (0, 2, 1)$, $\mathbf{y} = (5, 0, 1)$ and $\mathbf{y}' = (0, 0, 1)$. Suppose that the only strict rankings of $\succsim_1$ are $\mathbf{x} \succ_1 \mathbf{x}' \succ_1 \mathbf{y}'$ whereas the only strict ranking of $\succsim_2$ is $\mathbf{x}' \succ_2 \mathbf{y}'$. It is verified that the vectors of revealed relevant attributes are $\mathbf{m}_1 = (1, 1, 0)$ and $\mathbf{m}_2 = (0, 1, 0)$ respectively. Observe that $\mathbf{x}' \circ \mathbf{m}_1 = \mathbf{x} \circ \mathbf{m}_2$, hence $\mathbf{x}$ and $\mathbf{x}'$ have the same relevant attributes at periods 1 and 2. Similarly, $\mathbf{y}' \circ \mathbf{m}_1 = \mathbf{y} \circ \mathbf{m}_2$. Therefore, this sequence of choices violate Restricted Reversals, given that $\mathbf{x}' \succ_1 \mathbf{y}'$ whereas $\mathbf{x} \sim_2 \mathbf{y}$.*

A consequence of this axiom is the existence of a bijection between vectors of revealed relevant attributes and preference relations. Namely, this axiom is necessary and sufficient to represent the sequence of preferences $(\succsim_t)_t$ by the sequence of revealed relevant attributes $(\mathbf{m}_t)_t$ together with a binary relation over subsets of attributes. Formally, for any period t, let $X(\mathbf{m}_t) = \{\mathbf{x} \circ \mathbf{m}_t : \mathbf{x} \in X\}$ be the set of alternatives "filtered" through the revealed relevant attributes $\mathbf{m}_t$, and denote $\bar{X} = \bigcup_t X(\mathbf{m}_t)$.

PROPOSITION 16. *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals if and only if there exists a complete binary relation $\geqslant$ (called the **attribute ordering**), such that for any period t and any $\mathbf{x}, \mathbf{y} \in X$:*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t. \tag{3.1}$$

The interpretation is that DM has a fundamental preference—called the *attribute ordering*—that, unlike her choices $(\succsim_t)_t$, does not change over time. This

attribute ordering ranks vectors of attributes and does not depend on the relevant attributes.[5] The main consequence of Proposition 16 is that preference change can only be induced by changes in relevant attributes. Observe that the attribute ordering need not be transitive. We derive necessary and sufficient conditions for a transitive attribute ordering in Section 3.1.6.

3.1.4 Principle of Deliberation

The second normative principle that guides our analysis is a principle of deliberation: DM must evaluate all possible preferences at time t and consistently choose the best feasible one according to some criterion. This translates into an acyclicity axiom, which states that if DM changes her preference once, every future change should be due to the discovery of some new attributes—i.e that were not involved in the first change.

ACYCLICITY. *Preferences $(\succsim_t)_t$ satisfy Acyclicity if for any t and any $t' > t+1$, if $\mathbf{m}_{t+1} \neq \mathbf{m}_{t'}$, then there exists k such that $m^k_{t'} \neq m^k_{t+1} = m^k_t$.*

Note that, as soon as several choice reversals are observed, the principle of deliberation implies the existence of a constraint on preference change. Indeed, would preference change be unconstrained, DM would directly reach her most preferred preference once and for all. We interpret this constraint as DM's awareness: she can change only the attributes she is aware of, that is, the ones she is able to question.

Example 1 (continued). *Suppose that $\mathbf{m}_1 = (0,0,1)$ and $\mathbf{m}_2 = (0,1,0)$, namely the recruiter makes gender relevant but experience irrelevant at the second period. This could be because on the market men are more experienced, implying a form of statistical discrimination. Therefore, DM must have been able to modify her relevant attributes (at least) on these two attributes. Acyclicity implies that she could never choose the following relevant attributes in the future: $(0,0,0)$ and $(0,1,1)$ as they were accessible between period 1 and period 2. Since she did not change the relevance of the education*

[5] In a slightly different framework, Dietrich and List (2013a) provide an equivalence result between this separability condition (their axiom 2) and the existence of an attribute ordering.

attribute, we conclude that she was not aware of this attribute at this point. Assuming for instance that education provides a fair criterion to rank the candidates, she could later on decide to remove again gender only if education is made relevant jointly, reaching $\mathbf{m}_3 = (1, 0, 0)$.

3.1.5 The Representation

The constraint on preference change in the representation is formalized by a sequence of vectors $(\mathbf{a}_t)_{t=1}^{T-1}$, which represents DM's **awareness** between each period t and $t+1$. Namely, $\mathbf{a}_t \subseteq \{0,1\}^K$ for any t and codes as 1 attributes that DM can modify and as 0 the ones that she cannot modify between t and $t+1$. An awareness vector $\mathbf{a} \in \{0,1\}^K$ together with a vector of relevant attributes $\mathbf{m} \in \{0,1\}^K$ defines a set of **reachable attributes** for the next period $R(\mathbf{m}, \mathbf{a})$:

$$R(\mathbf{m}, \mathbf{a}) \equiv \left\{ \mathbf{m}' \in \{0,1\}^K : \text{for all } k, \mathbf{a}^k = 0 \text{ implies } \mathbf{m}'^k = \mathbf{m}^k \right\}.$$

To state our main result, define for any set A and any linear order $P \subset A^2$, $\max(A, P) = \{a \in A \mid aPb,\ \forall b \in A\}$.

THEOREM 3 (Representation). *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals and Acyclicity if and only if there exists a complete binary relation $\geqslant$, a sequence of awareness $(\mathbf{a}_t)_t$ (with $\mathbf{a}_t \in \{0,1\}^K$), and a linear order $\rhd$,[6] such that, for any t and any $\mathbf{x}, \mathbf{y} \in X$,*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t \tag{3.1}$$

$$\{\mathbf{m}_{t+1}\} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd). \tag{3.2}$$

The principle of sufficient reason together with the principle of deliberation are necessary and sufficient for what we name a **deliberate preference change model**. If the tuple $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ satisfy the conditions in theorem 3, we say that it **rationalizes** $(\succsim_t)_t$. In this model, DM's behavior is represented by the maximization of two binary relations: a preference relation on alternatives that

[6] It is observationally equivalent to construct a linear order or a complete preorder together with a tie-breaking rule for the meta-choice such that if $\mathbf{m}_t = \mathbf{m}$ and $\mathbf{m}_{t'} \neq \mathbf{m}$ for some $t' > t$, then $\mathbf{m}_\tau \neq \mathbf{m}$ for all $\tau > t'$.

together with the relevant attributes determine choices in each period (3.1) and a meta-preference relation on vectors of relevant attributes that determine the change of preference between periods (3.2). The revealed preference implication of our model is that when we observe choice reversals between alternatives x and y, we should observe other choice reversals on alternatives that share attributes with x and y. For instance if an employer stops discriminating at work this should impact her preferences in other contexts, such as her political preferences.

The fact that attributes can only be made relevant or irrelevant—and that DM cannot change her "taste" (attribute ordering) towards an attribute due to the stability of the attribute ordering—might seem arbitrary at first but it is important for two reasons. First, it is essential for the testability of the model, as otherwise almost any sequence of observed choice behavior could be rationalized by changing DM's tastes. Second, if the space of attributes is correctly specified from the beginning, there is no need to change DM's tastes. For instance, if the employer makes "gender" irrelevant to avoid discrimination, but makes it relevant again in the future due to an affirmative action policy, this policy should be thought of an attribute that is complementary with the attribute "gender". Therefore, it is not that DM changes her tastes toward the attribute "gender", but that the combination of "gender" and "affirmative action" is strictly preferred to "gender" alone. This suggests that the specification of the attributes is a crucial step that the researcher should discuss carefully, and commit to before observing choice data to avoid ex-post rationalization.

What can be inferred if one of the two axioms is violated? First, a violation of Restricted Reversal indicates that preference changes do not arise from changes in DM's revealed relevant attributes. Indeed, it is a necessary and sufficient condition for the existence of a time-independent attribute ordering that rationalizes each period's preference together with a set of relevant attributes (see proposition 16). Therefore, the analyst's knowledge of what determines DM's preference is incomplete: we may not observe all attributes, or the attribute ordering may change because DM discovers new consequences of an attribute for instance. Second, a violation of Acyclicity suggests that DM does not change her preferences *rationally*, meaning that no linear order can ratio-

nalize the sequence of meta choices. Canonical examples of non-deliberate preference changes are nudges, conformism or random utility. Alternatively, a violation of these axioms may suggest that the revealed relevant attributes are not the "truly" relevant attributes for DM. In this case, DM's behavior could be rationalized by our model with a different sequence $(\mathbf{m}'_t)_t$.[7]

Finally, we emphasize that our model is complementary with Bayesianism to explain preference change. Even though evidence suggests that agents do not always follow Bayes' rule, we do not think that an exhaustive theory of social interactions could do without belief updating. Instead, we argue that preference change and belief updating can occur simultaneously. This thesis receives empirical support in experiments on fake news by Barrera et al. (2020) (cited in the introduction).

Regarding the uniqueness of the ingredients of a deliberate preference change model, without further restrictions, only the preferences (i.e the attribute ordering) and the meta-preferences are identified up to an arbitrary completion on irrelevant attributes. Neither the relevant attributes nor the awareness are uniquely identified in general.

THEOREM 4 (Uniqueness). *Let* $(\succcurlyeq, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *and* $(\succcurlyeq', \rhd', \mathbf{m}_t, \mathbf{a}'_t)$ *rationalize* $(\succsim_t)_t$. *We have the following property: any completion of* $\succcurlyeq \cap \succcurlyeq', \rhd \cap \rhd'$, *together with* $(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ *also rationalize* $(\succsim_t)_t$.

In Section 3.1.7, we derive sufficient conditions for the identification of the relevant attributes, in the case where the the preferences at each period are transitive as well as the attribute ordering. Hence we first study in the next section the characterization of deliberate preference changes with transitive attribute ordering.

3.1.6 Transitive Attribute Ordering

Our main representation theorem does not guarantee that the attribute ordering is transitive and does not require that the observed preferences $(\succsim_t)_t$ are

[7] Note that if one does not want to restrict attention to revealed relevant attributes, it is possible to write axioms on multiple "candidate" sequences of relevant attributes (details available upon request).

transitive. Indeed Restricted Reversals constraints choices only between pairs of periods which is not enough to guarantee transitivity. For instance, suppose that $\mathbf{x}, \mathbf{y} \in X(\mathbf{m}_t)$, $\mathbf{y}, \mathbf{z} \in X(\mathbf{m}_{t'})$ and $\mathbf{x}, \mathbf{z} \in X(\mathbf{m}_{t''})$ but $\mathbf{z} \notin X(\mathbf{m}_t), \mathbf{x} \notin X(\mathbf{m}_{t'})$ and $\mathbf{y} \notin X(\mathbf{m}_{t''})$. It could be that $\mathbf{x} \succ_t \mathbf{y}, \mathbf{y} \succ_{t'} \mathbf{z}$ and $\mathbf{z} \succ_{t''} \mathbf{x}$ because Restricted Reversals does not constraint choices on triplets of periods. In fact, this problem is more general and may arise with any number of periods strictly greater than two.

Transitivity of preferences is sometimes viewed as a condition for rationality, hence it might be of interest to characterize transitivity of the attribute ordering. The following axiom extends Restricted Reversals to address this problem.

STRONG RESTRICTED REVERSALS. *For any* $\{t_1, \ldots, t_n\}$ *and any* $\{\mathbf{x}_k, \mathbf{x}'_k\}_{k=1,\ldots,n}$ *such that, for* $k = 1, \ldots, n-1$,

$$\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}} \quad \textit{and} \quad \mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1},$$

preferences $(\succsim_t)_t$ *satisfy Strong Restricted Reversals if:*

$$\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k, \ \textit{for every } k = 1, \ldots, n-1 \implies \mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n.$$

PROPOSITION 17. *Suppose that preferences* $(\succsim_t)_t$ *are transitive. Preferences satisfy Strong Restricted Reversals and Acyclicity if and only if there exists a deliberate preference change model* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *that rationalizes them with* $\geqslant$ *being a complete preorder.*

3.1.7 Identification of the Revealed Relevant Attributes

The relevant attributes are typically not identified without further restrictions on preferences. This is the case because when we observe an indifference, we cannot always identify whether this is due to an attribute being irrelevant, or whether DM is indifferent towards this attribute in the attribute ordering.

Denote $\mathcal{M}(\succsim_t) = \{\mathbf{m} : \exists \text{ a preorder } \geqslant \text{ s.t. } (\mathbf{m}, \geqslant) \text{ rationalizes } \succsim_t\}$ the set of relevant attributes that rationalize preferences at t using a transitive attribute

ordering. We show that, under the assumption that the observed preferences $\succsim_t$ are transitive, the set of vectors of relevant attributes $\mathbf{m}$ that can be used to rationalize preferences in the baseline model has a lattice structure. The most parsimonious vector is the vector of revealed relevant attributes $\mathbf{m}_t$,[8] but in principle other vectors could be used to rationalize DM's preferences.

PROPOSITION 18. *Suppose that preferences $\succsim_t$ are transitive. If Restricted Reversal is satisfied, $\mathcal{M}(\succsim_t)$ is a lattice ordered by $\geq$. Its minimum is the vector of revealed relevant attributes $\mathbf{m}_t$ and its maximum is $(1, \ldots, 1)$.*

This indeterminacy problem between irrelevant attributes and indifference can be solved if we impose that indifference are *only* caused by an attribute being irrelevant. In this case, an indifference $\mathbf{x} \sim_t \mathbf{y}$ has a clear interpretation in the sense that there is no attribute that motivates DM to choose $\mathbf{x}$ over $\mathbf{y}$. This is the content of the following axiom.

JUSTIFIED INDIFFERENCE. *Preferences $(\succsim_t)_t$ satisfy Justified Indifference if for any t and any alternatives $\mathbf{x}, \mathbf{y} \in X$,*

$$\mathbf{x} \sim_t \mathbf{y} \implies |\mathbf{x} - \mathbf{y}| \circ \mathbf{m}_t = (0, \ldots, 0).$$

When Justified Indifference is satisfied and if we restrict attention to strict attribute ordering, the relevant attributes are uniquely identified by the revealed relevant ones. Formally, let $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m} : \exists \text{ a partial order } > \text{ s.t. } (\mathbf{m}, >) \text{ rationalizes } \succsim_t\}$ be the set of relevant attributes that rationalize preferences at t using a strict attribute ordering. When Justified Indifference is satisfied, we have $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.

THEOREM 5. *Suppose that preferences $(\succsim_t)_t$ are transitive. Preferences satisfy Strong Restricted Reversal, Acyclicity and Justified Indifference if and only if there exists a deliberate preference change model $(>, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ that rationalizes $(\succsim_t)_t$ with $>$ being a partial order. Furhermore, for any period t, $\mathcal{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.*

[8] If X is not rich, the vector of revealed relevant attributes $\mathbf{m}_t$ need not be the minimum of the lattice.

3.2 Motivated Preference Change

Our main representation theorem shows that preference change can be represented by the maximization of a meta-preference. The representation, however, does not provide a straightforward interpretation of the meta-preference. It could be that DM is changing her behavior to make it more aligned with her values, or she may change preferences to serve her own-interests instead of purely disinterested motives—this is referred to as *motivated preference change*. In this section, we investigate the latter idea. We show that motivated preference change admits a tractable functional representation—this proves convenient for applications in the next section.

First, we construct an extension of the attribute ordering which allows us to keep track of (i) preferences over perceived alternatives at period t, and (ii) preferences over perceived alternatives at period t if she were to change her preferences to make good alternatives even better.

DEFINITION 2. *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$. *Denote* $\mathbf{a} \gg_t \mathbf{b}$ *if* $\mathbf{x} \circ \mathbf{m}_t = \mathbf{a}$ *for some* $\mathbf{x} \in X$ *and*

(i) $\mathbf{y} \circ \mathbf{m}_t = \mathbf{b}$ *for some* $\mathbf{y} \in X$ *and* $\mathbf{x} \succsim_t \mathbf{y}$; *or*

(ii) $\mathbf{y} \circ \mathbf{m} = \mathbf{b}$ *for some* $\mathbf{y} \in X$, $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, *and* $\mathbf{x} \succsim_t \mathbf{z}$ *for all* $\mathbf{z} \in X$.

The following axiom, which extends Strong Restricted Reversals, guarantees that DM makes attributes relevant if and only if these attributes are valued positively—that is, making these attributes (ir)relevant increases DM's utility.

MOTIVATED RESTRICTED REVERSALS. *Preferences* $(\succsim_t)_t$ *satisfy Motivated Restricted Reversals if for any* $\{t_1, \ldots, t_n\}$ *and any* $(\mathbf{a_k})_{k=1,\ldots,n} \in (\mathbb{R}^K)^n$ *such that* $\mathbf{a}_{k+1} \gg_{t_k} \mathbf{a}_k$ *for* $k = 1, \ldots, n-1$,

$$\mathbf{a}_1 \gg_{t_n} \mathbf{a}_n \implies \mathbf{a}_1 \ll_{t_n} \mathbf{a}_n.$$

The next axiom guarantees that there are no indifference between vectors of relevant attributes when changing preferences. Intuitively, the axiom states that if there is a tie between two vectors $\mathbf{m}$ and $\mathbf{m}'$ that yield identical utility,

DM breaks the tie in favor of one vector by virtually increasing her utility for some alternative $x \in X$ so that $\mathbf{m}$ becomes strictly preferred to $\mathbf{m}'$.

MOTIVATED TIE-BREAKING. *Preferences $(\succsim_t)_t$ satisfy Motivated Tie-Breaking if for all t, all $\mathbf{x} \in \max(X, \succsim_t)$, and all $\mathbf{y}, \mathbf{y}' \in X$ such that there exists $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$ with $\mathbf{y}' \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m} \circ \mathbf{m}_t$,*

$$\mathbf{y}' \in \max(X, \succsim_t) \implies \mathbf{m} = \mathbf{m}_t.$$

These two axioms are necessary and sufficient for the motivated preference change representation.

THEOREM 6 (Representation)**.** *Suppose that preferences $(\succsim_t)_t$ are transitive. Preferences $(\succsim_t)_t$ satisfy Motivated Restricted Reversals and Motivated Tie-Breaking if and only if there exists a sequence of awareness $(\mathbf{a}_t)_t$ and a function $u : \mathbb{R}^K \times \{0,1\}^K \longrightarrow \mathbb{R}$ such that for all t, all $\mathbf{m}_t$ and all $\mathbf{x}$,*

$$\mathbf{x} \succsim_t \mathbf{x}' \iff u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{x}' \circ \mathbf{m}_t)$$

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\operatorname{argmax}} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}).$$

As in the previous representation, DM chooses alternatives to maximize her attribute ordering, which can be represented by a utility function here. The main difference is that preference change must maximize DM's utility. Therefore, all attributes that are "negatively valued" will be made irrelevant as soon as possible, and all attributes that are "positively valued" will be made relevant as soon as possible.

3.3 An Application

An important feature of the model is path dependence—that is, the order in which DM becomes aware of certain attributes has a strong impact on the path of preference change. We illustrate this aspect in a voting context: ex-ante identical voters deliberately ignore what other voters think is relevant later on be-

cause this would undermine their view of their preferred candidate.[9] Therefore, we show that our model can account for polarization of political preferences among ex-ante identical voters in a simple and intuitive way.

Polarization refers to disagreement on policy issues or distrust of the other party members among politicians and citizens (Iyengar et al., 2019). There is now widespread agreement concerning the growing importance of ideological divisions both among politicized and educated voters as well as non-politicized citizens (Abramowitz and Saunders, 2008). There is no agreement, however, on the causes of polarization.[10]

From a Bayesian perspective, it is surprising that polarization increases as rational agents whose posterior beliefs are common knowledge cannot agree to disagree, even if their posteriors are based on different observed information about the world (Aumann, 1976). Arguing that voters have different priors certainly explains polarization, but it only moves the goalpost: where do differences in prior come from? Instead, our model provides a foundation for the concept of "partisan social identity" introduced in the political science literature (Iyengar and Westwood, 2015). This theory captures the tendency of voters to classify opposing partisans as members of an outgroup and copartisans as members of an ingroup. We show that our model can account for the construction of such opposing groups, and how partisan cues can reinforce division.

We consider a very stylized model with motivated preference change. There are two voters i and j and two candidates: $\mathbf{x}^D = (x^1, x^2, x^3)$ and $\mathbf{x}^R = (\tilde{x}^1, \tilde{x}^2, \tilde{x}^3)$ with $\tilde{x}^1 < 0 < x^1$, $x^2 < 0 < \tilde{x}^2$, $x^3 < 0 < \tilde{x}^3$ and $\tilde{x}^2 - \tilde{x}^1 > x^1 - x^2$. The first attribute captures the candidates' support for social policies (e.g. health care), the second attribute captures how conservative candidates are, and the third attribute represents corruption. Voters are ex-ante identical: they both value integrity and prefer candidates with strong convictions (represented by a high absolute value of the difference between the first and the second attributes).

[9] Note that Bayesian updating cannot induce this type of path dependence because it is order invariant (Cripps, 2018).

[10] Recent finding suggests that the emergence of the internet or rising economic inequality are less plausible causes than changes that are specific to the US (e.g., changing party composition, growing racial divisions, or the emergence of partisan cable news) (Boxell et al., 2020).

We can represent their preferences as follows:

$$u(\mathbf{x} \circ \mathbf{m}) = \left(x^1 m^1 - x^2 m^2\right)^2 - x^3 m^3.$$

Suppose that voter i attends a political debate with both candidates: $\mathbf{a}_1^i = (1, 1, 0)$. She will change her preferences and value more candidate $\mathbf{x}^R$ who has stronger convictions: the meta-maximization writes

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)) = \left(\tilde{x}^2 - \tilde{x}^1\right)^2 &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,1,0)) = \left(\tilde{x}^2\right)^2 \\ &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,0,0)) = \left(x^1\right)^2 \\ &> 0 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,0)).\end{aligned}$$

Later, voter i becomes aware that candidate $\mathbf{x}^R$ is corrupted: $\mathbf{a}_2^i = (0, 0, 1)$. She decides to ignore this information and keep this attribute irrelevant if:

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,1)) &= \max\left\{\left(\tilde{x}^2 - \tilde{x}^1\right)^2 - \tilde{x}^3, \left(x^1 - x^2\right)^2 - x^3\right\} \\ &< \left(\tilde{x}^2 - \tilde{x}^1\right)^2 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)).\end{aligned}$$

i.e. whenever $\left(\tilde{x}^2 - \tilde{x}^1\right)^2 > \left(x^1 - x^2\right)^2 - x^3$. Namely, whenever candidate $\mathbf{x}^R$ has strong convictions that counterbalance her corruption. The intuition is that making "corruption" relevant would undermine her view of candidate $\mathbf{x}^R$. In the end, voter i's most preferred candidate is $\mathbf{x}^R$.

Instead, voter j first becomes aware of a felony committed by candidate $\mathbf{x}^R$: $\mathbf{a}_1^j = (0, 0, 1)$. She will change her preferences to make it relevant: the meta-maximization writes

$$\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,1)) = -x^3 > 0 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,0)).$$

At this point voter j prefers the upstanding candidate $\mathbf{x}^D$.

Later, voter j attends a political debate with both candidates: $\mathbf{a}_2^j = (1, 1, 0)$. She will lean toward the candidate $\mathbf{x}^D$ even though he has less convictions than the candidate $\mathbf{x}^R$ whenever $(x^1 - x^2)^2 - x^3 > (\tilde{x}^2 - \tilde{x}^1)^2 - \tilde{x}^3$. Namely, whenever the convictions of $\mathbf{x}^R$ does not make up for his felonies. In the end, voter j's

most preferred candidate is $\mathbf{x}^D$.

It is quite striking that two identical voters who become aware of the same attributes can become polarized. This arises due to the path dependence of preference change: past justifications can conflict with new justifications leading to rich dynamics.

Proofs

Proof of Proposition 16. We say that a pair $(\mathbf{m}, \geqslant)$ **represents** $\succsim$ if for any $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \succsim \mathbf{y} \iff \mathbf{x} \circ \mathbf{m} \geqslant \mathbf{y} \circ \mathbf{m}$.

(*Necessity*). Suppose there exists $\geqslant$ such that for every t, $(\mathbf{m}_t, \geqslant)$ **represents** $\succsim_t$. Let $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$ four alternatives such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'} =: \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'} =: \mathbf{b}$ for some periods t, t'. Therefore, $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{a} \geqslant \mathbf{b} \iff \mathbf{x}' \succsim_{t'} \mathbf{y}'$.

(*Sufficiency*). Suppose that $(\succsim_t)_t$ satisfy Restricted Reversals. Let $X(\mathbf{m}_t) = \{\mathbf{x} \circ \mathbf{m}_t : \mathbf{x} \in X\}$. First, we fix a period t and show that we can indeed construct an ordering $\geqslant_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \geqslant_t)$ represents $\succsim_t$. We define the two following binary relations on $X(\mathbf{m}_t)$:

$$>_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \succ_t \mathbf{y}\},$$
$$\simeq_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \sim_t \mathbf{y}\}.$$

By definition, $\simeq_t$ is reflexive and symmetric. We show that $>_t$ is irreflexive, i.e. for any $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \neq \mathbf{y}$ and $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$, $\mathbf{x} \sim_t \mathbf{y}$. Suppose by contradiction that (w.l.o.g) $\mathbf{x} \succ_t \mathbf{y}$ and denote M' the set of attributes on which $\mathbf{x}$ and $\mathbf{y}$ differ. Given that M' is not revealed relevant, there must exist some alternatives $\mathbf{x}'$ and $\mathbf{y}'$ that differ only on a strict subset $M'' \subset M'$ and are strictly ranked according to $\succsim_t$. Therefore, M'' should be revealed relevant, a contradiction.

Now let $\mathbf{a}, \mathbf{b} \in X(\mathbf{m}_t)$, with $\mathbf{a} \neq \mathbf{b}$, and $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_t = \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_t = \mathbf{b}$. Applying Restricted Reversal with $t = t'$, we obtain $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_t \mathbf{y}'$. Given that $>_t$ is irreflexive, this establishes that it is asymmetric. It also proves that $>_t \cap \simeq_t = \emptyset$.

Therefore, by defining $\succcurlyeq_t := \simeq_t \cup >_t$, $\succcurlyeq_t$ is complete on $X^2(\mathbf{m}_t)$ (by the completeness of $\succsim_t$) and reflexive, $\simeq_t$ and $>_t$ respectively are its symmetric and asymmetric parts. Furthermore, $(\mathbf{m}_t, \succcurlyeq_t)$ represents $\succsim_t$.

Second, we show that for any two distinct periods t and t', $\succcurlyeq_t$ does not contradict $\succcurlyeq_{t'}$. Let $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}'$ be such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'} =: \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'} =: \mathbf{b}$. Then

$$\begin{aligned} \mathbf{x} \succsim_t \mathbf{y} &\iff \mathbf{x} \circ \mathbf{m}_t \succcurlyeq_t \mathbf{y} \circ \mathbf{m}_t \\ &\iff \mathbf{a} \succcurlyeq_t \mathbf{b}; \\ \mathbf{x}' \succsim_{t'} \mathbf{y}' &\iff \mathbf{x}' \circ \mathbf{m}_{t'} \succcurlyeq_{t'} \mathbf{y}' \circ \mathbf{m}_{t'} \\ &\iff \mathbf{a} \succcurlyeq_{t'} \mathbf{b}. \end{aligned}$$

Restricted Reversals implies that $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_{t'} \mathbf{y}'$. Hence we conclude that:

$$\mathbf{a} \succcurlyeq_t \mathbf{b} \iff \mathbf{a} \succcurlyeq_{t'} \mathbf{b}.$$

Finally, we define $\succcurlyeq := \bigcup_t \succcurlyeq_t$. By the previous argument, $\succcurlyeq \cap X^2(\mathbf{m}_t) = \succcurlyeq_t$, so for any t $(\mathbf{m}_t, \succcurlyeq)$ represents $\succsim_t$. Furthermore $\succcurlyeq$ is complete on $\bar{X}$. □

Proof of Theorem 3. (Necessity). We prove the necessity of Acyclicity. Let t be a given period. We denote $t' \equiv \min\{\tau > t+1 : \mathbf{m}_\tau \neq \mathbf{m}_{t+1}\}$ and suppose that it is well defined —if it is not so, Acyclicity is trivially satisfied. Fix $\tau \geq t'$ and suppose that $|\mathbf{m}_\tau - \mathbf{m}_t| \leq |\mathbf{m}_{t+1} - \mathbf{m}_t|$. The constraint of awareness implies that $|\mathbf{m}_{t+1} - \mathbf{m}_t| \leq \mathbf{a}_t$, hence $|\mathbf{m}_\tau - \mathbf{m}_t| \leq \mathbf{a}_t$, i.e. $\mathbf{m}_\tau \in R(\mathbf{m}_t, \mathbf{a}_t)$. But, by the transitivity of $\rhd$, we have $\mathbf{m}_\tau \rhd \mathbf{m}_{t+1}$ and thus $\mathbf{m}_{t+1} \neq \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd)$, a contradiction. Hence Acyclicity is satisfied. We already proved the necessity of Restricted Reversal from proposition 16.

(Sufficiency). We know from proposition 16 that there exists an attribute ordering $\succcurlyeq \subseteq \bar{X}^2$, such that for any period t, $(\mathbf{m}_t, \succcurlyeq)$ represent $\succsim_t$. We first construct a sequence $\mathbf{a}_t$. For any t, let $\mathbf{a}_t = |\mathbf{m}_{t+1} - \mathbf{m}_t|$. Hence the set of reachable relevant attributes reduces to:

$$R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t+1}|) = \{\mathbf{m} : \mathbf{m}_t \wedge \mathbf{m}_{t+1} \leq \mathbf{m} \leq \mathbf{m}_t \vee \mathbf{m}_{t+1}\}.$$

where $\wedge$ and $\vee$ are the element-wise minimum and maximum, respectively. Define the revealed meta-preference relation $\rhd$ as follows: $\mathbf{m} \rhd \mathbf{m}'$ if $\mathbf{m} \neq \mathbf{m}'$ and there exists t, such that $\mathbf{m} = \mathbf{m}_t$ and,

$$\mathbf{m}' \in \bigcup_{t' : t' < t} R(\mathbf{m}_{t'}, \mathbf{a}_{t'}).$$

We verify that $\rhd$ is asymmetric. Suppose that $\mathbf{m} \rhd \mathbf{m}'$ and take $t' < t$, such that $\mathbf{m}' \in R(\mathbf{m}_{t'}, \mathbf{a}_{t'})$. Let us first show that Acyclicity implies that there cannot be any $t'' > t$ such that $\mathbf{m}' = \mathbf{m}_{t''}$. Assume by contradiction that such a t'' exists. Then, we have

$$|\mathbf{m}_{t''} - \mathbf{m}_{t'}| \underbrace{=}_{\text{Def. } \mathbf{m}_{t''}} |\mathbf{m} - \mathbf{m}_{t'}| \underbrace{\leq}_{\mathbf{m}' \in R(\mathbf{m}_{t'}, |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|)} |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|$$

Which, by Acyclicity, implies that $\mathbf{m}_{t'+1} = \mathbf{m}_{t''} = \mathbf{m}$. But, then we still have that $\mathbf{m} \in R(\mathbf{m}_{t'+1}, |\mathbf{m}_{t'+2} - \mathbf{m}_{t'+1}|)$ so that, applying the previous reasoning inductively, we obtain $\mathbf{m}' = \mathbf{m}_{t'+2} = \mathbf{m}_{t'+3} = \dots \mathbf{m}_t = \mathbf{m} \neq \mathbf{m}'$. A contradiction. Second assume by contradiction that $\mathbf{m} = \mathbf{m}_{t'''}$ and $\mathbf{m} \in R(\mathbf{m}_{t''}, |\mathbf{m}_{t''+1} - \mathbf{m}_{t''}|)$ for some t'', t''' such that $t'' < t''' < t$. By the same argument, Acyclicity would then imply that $\mathbf{m} = \mathbf{m}_{t''+2} = \mathbf{m}_{t''+3} = \dots \mathbf{m}_{t'''} = \mathbf{m}' \neq \mathbf{m}$. A contradiction.

We now verify that $\rhd$ is transitive. Suppose that $\mathbf{m} \rhd \mathbf{m}'$ and $\mathbf{m}' \rhd \mathbf{m}''$. Then there exist t, t' with $t > t'$, such that, $\mathbf{m} = \mathbf{m}_t$ and $\mathbf{m}' = \mathbf{m}_{t'}$. Moreover,

$$\mathbf{m}'' \in \bigcup_{t''' : t''' < t'} R(\mathbf{m}_{t'''}, \mathbf{a}_{t'''}) \subseteq \bigcup_{t'' : t'' < t} R(\mathbf{m}_{t''}, \mathbf{a}_{t''})$$

where the inclusion follows from $t > t'$. We conclude that $\mathbf{m} \rhd \mathbf{m}''$, implying the transitivity of $\rhd$.

Furthermore, by the definition of $\rhd$, $\mathbf{m}_{t+1} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd)$. By Szpilrajn's theorem, the meta-preference can be completed on vectors that are not ranked yet.

□

Proof of Theorem 4. The fact that, for any period t, any completion of $\geqslant \cap \geqslant'$, together with $\mathbf{m}_t$, represent $\succsim_t$ follows directly from the proof of Proposition

18.

We next show that by considering $\rhd \cap \rhd'$ and the sequence of awareness $(\mathbf{a}_t \circ \mathbf{a}'_t)_t$, we can rationalize the meta-choices of each period t. Fix a period t, and suppose that being at $\mathbf{m}_t$, DM faces the meta-menu $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$. Note that $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t) = R(\mathbf{m}_t, \mathbf{a}_t) \cap R(\mathbf{m}_t, \mathbf{a}'_t)$. Hence it implies that $(\mathbf{m}_{t+1}, \mathbf{m}) \in \rhd \cap \rhd'$ for any $\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$.

This completes the proof than any completion of $\geqslant \cap \geqslant'$ and $\rhd \cap \rhd'$, together with $(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ rationalize $(\succsim_t)_t$. □

Proof of Proposition 17. (*Necessity.*) Suppose there exists a complete preorder $\geqslant$ such that for every t, $(\mathbf{m}_t, \geqslant)$ represents $\succsim_t$. Take any $\{t_1, \ldots, t_n\}$ and any $\{\mathbf{x_k}, \mathbf{x'_k}\}_{k=1,\ldots,n}$ such that, for $k = 1, \ldots, n-1$:

$$\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}},$$
$$\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1},$$

and for every $k \leq n-1$, $\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k$. The latter implies that $\mathbf{x}_k \circ \mathbf{m}_{t_k} \geqslant \mathbf{x}'_k \circ \mathbf{m}_{t_k}$. Hence by the transitivity of $\geqslant$, we can conclude that $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1} \geqslant \mathbf{x}_{n-1} \circ \mathbf{m}_{t_{n-1}} = \mathbf{x}_n \circ \mathbf{m}_{t_n}$, i.e. $\mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n$. Hence Strong Restricted Reversal is satisfied.

(*Sufficiency.*) We fix a period t and we show that we can construct a complete preorder $\geqslant_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \geqslant_t)$ represents $\succsim_t$. We define $\geqslant_t$ in the same way as in the proof of proposition 16. Given that Strong Restricted Reversal implies Restricted Reversal, the same arguments apply and we conclude that $(\mathbf{m}_t, \geqslant_t)$ represents $\succsim_t$. Furthermore, the transitivity of $\geqslant_t$ is a direct consequence of the transitivity of $\succsim_t$. We need now to construct a complete preorder $\geqslant$ that is time-independent.

From the proof of proposition 16, we know that for any two distinct periods t and t', $\geqslant_t$ does not contradict $\geqslant_{t'}$. We define $\geqslant_{1;T} := \bigcup_t \geqslant_t$. We know therefore that for any t, $\geqslant_{1;T} \cap X^2(\mathbf{m}_t) = \geqslant_t$.

We next show that the transitive closure of $\geqslant_{1;T}$, denoted $\geqslant^C_{1;T}$, can represent the sequence $(\succsim_t)_t$ together with the sequence $(\mathbf{m}_t)_t$. Namely, we show that for any period t, any $\mathbf{a}, \mathbf{b} \in X(\mathbf{m}_t)$, if $\mathbf{b} >_t \mathbf{a}$, there cannot be a sequence $(\mathbf{a}_k)_{k=1,\ldots,n}$

in $\bar{X}^2$ such that $\mathbf{a}_1 = \mathbf{a}$, $\mathbf{a}_n = \mathbf{b}$, and for any $k \leq n-1$, there exists t_k such that $\mathbf{a}_k \succcurlyeq_{t_k} \mathbf{a}_{k+1}$. Let suppose by contradiction the existence of such a sequence. If $t_1 \neq t$, then complete the sequence with $\mathbf{a}_0 = \mathbf{a}$ and $t_0 = t$; similarly, if $t_n \neq t$, then complete the sequence with $\mathbf{a}_{n+1} = \mathbf{b}$ and $t_n = t$. Therefore, w.l.o.g we consider the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$.

If $t_k = t_{k'}$ for some $k \neq k'$, we show that we can restrict to a subsequence $(\mathbf{a}_{\tau(k)})_{k=1,\dots,n+1}$ with $\tau(0) = 0$, $\tau(n+1) = n+1$, such that $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$. Let suppose that $t_k = t_{k'}$ with $k < k'$ and that for any $k \leq i, j < k'$, if $i \neq j$ then $t_i \neq t_j$. Let's consider the sequence $(\mathbf{a}_i)_{k \leq i \leq k'+1}$. There exists a sequence $(\mathbf{x}_i, \mathbf{y}_{i+1})_{k \leq i \leq k'}$ such that $\mathbf{x}_k \circ \mathbf{m}_{t_k} = \mathbf{a}_k$, $\mathbf{y}_{k'+1} \circ \mathbf{m}_{t_{k'}} = \mathbf{a}_{k'+1}$, for any $k \leq i \leq k'-1$, $\mathbf{y}_{i+1} \circ \mathbf{m}_{t_i} = \mathbf{x}_{i+1} \circ \mathbf{m}_{t_{i+1}} = \mathbf{a}_{i+1}$, and for any $k \leq i \leq k'$, $\mathbf{x}_i \succsim_{t_i} \mathbf{y}_{i+1}$. By applying Strong Restricted Reversal, this must be that $\mathbf{x}_k \succsim_{t_k} \mathbf{y}_{k'+1}$, i.e. $\mathbf{a}_k \succcurlyeq_{t_k} \mathbf{a}_{k'+1}$. Therefore, from the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$, we can construct a subsequence $(\mathbf{a}_{\tau(k)})_{k=0,\dots,n+1}$, with $\tau(0) = 0$, $\tau(n+1) = n+1$, $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$, and such that for any k with $\tau(k) \neq \tau(k+1)$, $\mathbf{a}_{\tau(k)} \succcurlyeq_{\tau(k)} \mathbf{a}_{\tau(k+1)}$. From a similar reasoning, we conclude by Strong Restricted Reversal that $\mathbf{a} \succcurlyeq_t \mathbf{b}$, a contradiction.

By an implication of Szpilrajn's theorem (see Corollary A.1 in Ok (2007)), there exists a complete, transitive and reflexive binary relation that extends $\succcurlyeq^C_{1;T}$. We denote it $\succcurlyeq$. We proved that for any t, $X^2(\mathbf{m}_t) \cap \succcurlyeq = \succcurlyeq_t$, hence $(\mathbf{m}_t, \succcurlyeq)$ represents $\succsim_t$. □

LEMMA 1. *Assume richness and the transitivity of $\succsim_t$ for any t. If $M \subset \{1, \dots, K\}$ is revealed relevant, then M is a singleton.*

Proof of Lemma 1. Let M be revealed relevant and suppose by contradiction that $|M| = n > 1$. This means that there exists $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}^{-M} = \mathbf{y}^{-M}$ and $x^k \neq y^k$ for any $k \in M$, with $\mathbf{x} \not\sim_t \mathbf{y}$; and for every $M' \subsetneq M$ and every $\mathbf{w}, \mathbf{z} \in X$ with $\mathbf{w}^{-M'} = \mathbf{z}^{-M'}$, $\mathbf{w} \sim_t \mathbf{z}$. By the richness assumption, there exists a sequence of alternatives $\mathbf{z}_1, \dots, \mathbf{z}_n \in X$ such that $\mathbf{z}_1 = \mathbf{x}$, $\mathbf{z}_n = \mathbf{y}$ and $z_i^{-k} = z_{i+1}^{-k}$ for some $k \in M$, for all $i = 1, \dots, n-1$. By assumption, it must be that $\mathbf{z}_i \sim_t \mathbf{z}_{i+1}$ for all $i = 1, \dots, n-1$, which by transitivity would imply that $\mathbf{x} \sim_t \mathbf{y}$, a contradiction. □

Proof of Proposition 18. First, we show that $(1, \dots, 1)$ can rationalize $\succsim_t$. In this

case, our representation at t coincides with standard preference maximization because for any $\mathbf{x} \in X$, $\mathbf{x} \circ (1, \dots, 1) = \mathbf{x}$. Identifying $\geqslant_t$ with $\succsim_t$ yields the desired result.

Second, we show that for any $\mathbf{m}' \not\geq \mathbf{m}_t$, $(\mathbf{m}', \geqslant'_t)$ cannot rationalize $\succsim_t$ for some $\geqslant'_t$. By contradiction, suppose that there exists such $\mathbf{m}'$. Given lemma 1 and the definition of $\mathbf{m}_t$, there exists an attribute k such that $m_t^k - m'^k = 1$, and some alternatives $\mathbf{x}, \mathbf{y}$ such that $\mathbf{x}^{-k} = \mathbf{y}^{-k}$, $x^k \neq y^k$ and $\mathbf{x} \not\succsim_t \mathbf{y}$ for some $\mathbf{x}, \mathbf{y} \in X$. Given that $\mathbf{x} \circ \mathbf{m}' = \mathbf{y} \circ \mathbf{m}'$, this contradicts the fact that $(\mathbf{m}', \geqslant'_t)$ rationalizes $\succsim_t$.

Finally, we prove that for any $\mathbf{m}' > \mathbf{m}_t$, there exists $\geqslant'_t$ such that $(\mathbf{m}', \geqslant'_t)$ rationalizes $\succsim_t$. Define:

$$
\begin{aligned}
>'_t &= \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}') \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}', \mathbf{b} = \mathbf{y} \circ \mathbf{m}', \text{ and } \mathbf{x} \succ_t \mathbf{y}\}, \\
\simeq'_t &= \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}') \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}', \mathbf{b} = \mathbf{y} \circ \mathbf{m}', \text{ and } \mathbf{x} \sim_t \mathbf{y}\}.
\end{aligned}
$$

A similar reasoning as in the proof of proposition 16 establishes that $(\mathbf{m}', \geqslant'_t)$ rationalizes $\succsim_t$. □

Proof of Theorem 5. The proof of the necessity of Justified Indifference is left to the readers. By proposition 18, $\mathbf{m}_t \in \mathcal{M}(\succsim_t)$, so that there exists $\geqslant$ such that for all $\mathbf{x}, \mathbf{y} \in X$

$$
\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t \tag{3}
$$

Moreover, the contraposition of Justified Indifference implies that for all $\mathbf{x}, \mathbf{y} \in X$, if $\mathbf{x} \circ \mathbf{m}_t \neq \mathbf{y} \circ \mathbf{m}_t$, then either $\mathbf{x} \succ_t \mathbf{y}$ or $\mathbf{x} \succ_t \mathbf{y}$. Hence, for any $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{x} \circ \mathbf{m}_t \neq \mathbf{y} \circ \mathbf{m}_t$, (3) implies that $\mathbf{x} \circ \mathbf{m}_t > \mathbf{y} \circ \mathbf{m}_t$. Hence, $\mathbf{m}_t \in \mathcal{M}^\star(\succsim_t)$.

Now assume by contradiction that there exists $\mathbf{m} \in \mathcal{M}^\star(\succsim_t)$ with $\mathbf{m} \neq \mathbf{m}_t$. Given that $\mathbf{m}_t$ is minimal this requires $\mathbf{m}$ to be such that $m_t^i = 1 \implies m^i = 1$. Given that $\mathbf{m} \neq \mathbf{m}_t$, we know there exists i such that $m_t^i = 0$ and $m^i = 1$. Our preliminary richness requirement about the set X combined with the richness assumption implies that there exist two alternatives $\mathbf{x}, \mathbf{y} \in X$ such that $x^i \neq y^i$ and $\mathbf{x}^{-i} = \mathbf{y}^{-i}$. This means that $\mathbf{x} \circ \mathbf{m} \neq \mathbf{y} \circ \mathbf{m}$. Given that $\mathbf{m} \in \mathcal{M}^\star(\succsim_t)$, this means that there exists $>$ such that either $\mathbf{x} \circ \mathbf{m} > \mathbf{y} \circ \mathbf{m}$ or $\mathbf{x} \circ \mathbf{m} < \mathbf{y} \circ \mathbf{m}$ that

rationalizes $\succsim_t$. Hence, we either have $\mathbf{x} \succ_t \mathbf{y}$ or $\mathbf{y} \succ_t \mathbf{x}$. Furthermore, $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$, which, given that $\mathbf{m}_t \in \mathcal{M}^\star(\succsim_t)$, implies that $\mathbf{x} \sim_t \mathbf{y}$, a contradiction. □

Proof of Theorem 6. (Necessity) Assume that there exists a sequence of awareness $(\mathbf{a}_t)_t$ and a function $u : \mathbb{R}^K \times \{0,1\}^K \longrightarrow \mathbb{R}$ such that for all t, all $\mathbf{m}_t$ and all $\mathbf{x}$,

$$\mathbf{x} \precsim_t \mathbf{x}' \iff u(\mathbf{x} \circ \mathbf{m}_t) \leq u(\mathbf{x}' \circ \mathbf{m}_t)$$

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_{t+1}, \mathbf{a}_t)}{\arg\max} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}).$$

Step 1: We show that for all t if $\mathbf{a} \gg_t \mathbf{b}$, then $u(\mathbf{a}) \geq u(\mathbf{b})$.

To see this, assume first that there exists $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{a} = \mathbf{x} \circ \mathbf{m}_t$ and $\mathbf{b} = \mathbf{y} \circ \mathbf{m}_t$. Then, $\mathbf{x} \succsim_t \mathbf{y}$ implies that $u(\mathbf{a}) = u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{y} \circ \mathbf{m}_t) = u(\mathbf{b})$. Now assume that there exists $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{a} = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{b} = \mathbf{y} \circ \mathbf{m}$ for some $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, and $\mathbf{x} \in \max(X, \precsim_t)$. Then $u(\mathbf{x} \circ \mathbf{m}_t) = \max_{\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m}) \geq u(\mathbf{y} \circ \mathbf{m}) = u(\mathbf{b})$.

Step 2. We show that MRR holds.

Suppose we have $\{t_1, \ldots, t_n\}$ and $\{\mathbf{a_k}\}_{k=1,\ldots,n} \in (\mathbb{R}^K)^n$ such that $\mathbf{a}_{k+1} \gg_{t_k} \mathbf{a}_k$ for $k = 1, \ldots, n-1$, and $\mathbf{a}_1 \gg_t \mathbf{a}_n$. Given *Step 1.*, this implies that

$$u(\mathbf{a}_n) \cdots \geq u(\mathbf{a}_2) \geq u(\mathbf{a}_1) \geq u(\mathbf{a}_n)$$

Hence, $u(\mathbf{a}_n) = u(\mathbf{a}_1)$. Moreover, from $\mathbf{a}_1 \gg_t \mathbf{a}_n$ we know that either there exists $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{a}_1 = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}_t$, and $\mathbf{x} \succsim_t \mathbf{y}$; or there exist $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{a}_1 = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}$ for some $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, and $\mathbf{x} \in \max(X, \succsim_t)$. If the first case holds, then since $u(\mathbf{a}_n) = u(\mathbf{a}_1)$, we have $\mathbf{x} \sim_t \mathbf{y}$ and, therefore, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}_t \gg_t \mathbf{x} \circ \mathbf{m}_t = \mathbf{a}_1$. If the second case holds then, if $\mathbf{m} = \mathbf{m}_t$ we are back to the first case; otherwise, if $\mathbf{m} \neq \mathbf{m}_t$, since $u(\mathbf{y} \circ \mathbf{m}) = u(\mathbf{a}_n) = u(\mathbf{a}_1) = u(\mathbf{y} \circ \mathbf{m}_t)$ and $\mathbf{m}_t \in \arg\max_{\mathbf{m} \in R(\mathbf{m}_{t+1}, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$, we have

$$\#\Big(\underset{\mathbf{m} \in R(\mathbf{m}_{t-1}, \mathbf{a}_t)}{\arg\max} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}) \Big) = 2,$$

which contradicts the fact that $\{\mathbf{m}_t\} = \arg\max_{\mathbf{m} \in R(\mathbf{m}_{t+1}, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$.

Step 3. We show that MTB holds.

Let t, $\mathbf{x} \in \max(X, \succsim_t)$, $\mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{y} \circ \mathbf{m} = \mathbf{y}' \circ \mathbf{m} \circ \mathbf{m}_t$ and $\mathbf{y}' \in \max(X, \precsim_t)$. We have that $u(\mathbf{y} \circ \mathbf{m}) = u(\mathbf{y}' \circ \mathbf{m} \circ \mathbf{m}_t)$. But then since $u(\mathbf{y}' \circ \mathbf{m} \circ \mathbf{m}_t) = \max_{\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m})$, if $\mathbf{m} \neq \mathbf{m}_t$, then

$$\#\Big(\arg\max_{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}) \Big) = 2,$$

which contradicts the fact that $\{\mathbf{m}_t\} = \arg\max_{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$.

(Sufficiency) Define $\geq^*$ as follows

$$\geq^* = \bigcup_{n \in \mathbb{N}} \bigcup_{t_1, t_2, \ldots, t_n, t_{n+1}} \{(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^2 : \exists (\mathbf{a}_k)_{k \leq n} \in \mathbb{R}^n, \mathbf{a} \gg_{t_n+1} \mathbf{a}_n \gg_{t_n} \ldots \mathbf{a}_1 \gg_{t_1} \mathbf{b}\}$$

Step 1: We show that for all $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t$.

First note that, by definition of $\gg_t$, $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \gg_t \mathbf{y} \circ \mathbf{m}_t$. Hence, we only need to prove that $\mathbf{x} \circ \mathbf{m}_t \gg_t \mathbf{y} \circ \mathbf{m}_t \iff \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t$. That $\mathbf{x} \circ \mathbf{m}_t \gg_t \mathbf{y} \circ \mathbf{m}_t \implies \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t$ directly follows from the definition of $\geq^*$. To show the converse, assume $\mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t$. This means that there exist $(n, (t_k)_{1 \leq k \leq n+1}, (\mathbf{a}_k)_{1 \leq k \leq n})$ such that

$$\mathbf{y} \circ \mathbf{m}_t \ll_{t_1} \mathbf{a}_1 \ll_{t_2} \cdots \ll_{t_{n+1}} \mathbf{x} \circ \mathbf{m}_t \tag{4}$$

Given that $(\mathbf{x} \circ \mathbf{m}_t, \mathbf{y} \circ \mathbf{m}_t) \in X^2(\mathbf{m}_t)$ and the completeness of $\precsim_t$ we either have that $\mathbf{x} \circ \mathbf{m}_t \gg_t \mathbf{y} \circ \mathbf{m}_t$ or $\mathbf{x} \circ \mathbf{m}_t \ll_t \mathbf{y} \circ \mathbf{m}_t$. If the former case holds there is nothing left to prove. If the later case holds, then, by MRR and (4), so is the former. Hence, $\mathbf{x} \circ \mathbf{m}_t \gg_t \mathbf{y} \circ \mathbf{m}_t \iff \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t$, as desired.

Step 2: We show that if $\mathbf{x} \in \max(X, \precsim_t)$, *then for all* $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|) \backslash \{\mathbf{m}_t\}$, *and all* $\mathbf{y} \in X$, *we have* $\mathbf{y} \circ \mathbf{m} <^* \mathbf{x} \circ \mathbf{m}_t$.

Assume that $\mathbf{x} \in \max(X, \precsim_t)$. Hence, for all $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$ and all $\mathbf{y} \in X$, $\mathbf{y} \circ \mathbf{m} \ll_t \mathbf{x} \circ \mathbf{m}_t$. This means that $\mathbf{y} \circ \mathbf{m}_t \leq^* \mathbf{x} \circ \mathbf{m}_t$. By contradiction,

suppose that for some $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|) \setminus \{\mathbf{m}_t\}$ and some $\mathbf{y} \in X$, $\mathbf{x} \circ \mathbf{m}_t \leq^* \mathbf{y} \circ \mathbf{m}$. This implies that there exists $(n, (t_k)_{1 \leq k \leq n+1}, (\mathbf{a}_k)_{1 \leq k \leq n})$ such that

$$\mathbf{x} \circ \mathbf{m}_t \ll_{t_1} \mathbf{a}_1 \ll_{t_2} \cdots \ll_{t_{n+1}} \mathbf{y} \circ \mathbf{m}.$$

From this and the fact that $\mathbf{y} \circ \mathbf{m} \ll_t \mathbf{x} \circ \mathbf{m}_t$, it follows from MRR that $\mathbf{x} \circ \mathbf{m}_t \ll_t \mathbf{y} \circ \mathbf{m}$. Hence by definition of $\ll_t$, there exists $\mathbf{y}' \in X$ such that $(\mathbf{y} \circ \mathbf{m}) \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_t$ and $\mathbf{y}' \succsim_t \mathbf{x}$. By MTB this implies that $\mathbf{m} = \mathbf{m}_t$, a contradiction.

Step 3: We conclude.

Now note that, by construction, $\geq^*$ is transitive and reflexive. Thus, it is a preorder. By an extension of Szpilrajn's theorem (see Corollary A.1 in Ok (2007)) we can complete $\geq^*$ to obtain a complete preorder. This means that there exists utility function u representing $\geq^*$. By *Step 1.*, we thus have that for all t and all $\mathbf{x}, \mathbf{y} \in X$,

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t \iff u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{y} \circ \mathbf{m}_t)$$

By *Step 2.* we have that for all t, all $\mathbf{x}, \mathbf{y} \in X$, and all $\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|) \setminus \{\mathbf{m}_t\}$

$$\mathbf{x} \in \max(X, \precsim_t) \implies \mathbf{x} \circ \mathbf{m}_t >^* \mathbf{y} \circ \mathbf{m} \iff u(\mathbf{x} \circ \mathbf{m}_t) > u(\mathbf{y} \circ \mathbf{m})$$

Hence, taking $\mathbf{a}_t = |\mathbf{m}_{t+1} - \mathbf{m}_t|$ for all t, we obtain,

$$\mathbf{m}_{t+1} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\arg\max} \max_{\mathbf{x} \in X} u(\mathbf{x}, \mathbf{m}).$$

□

Chapter 4

The Value of Model Misspecification in Communication[1]

Why are people drawn to monocausal explanations of complex social phenomena? One of the distinctive features of the increasingly popular populist narratives is their simplicity: many complex problems boil down to a unique explanation, such as immigration, the welfare state, or bureaucracy.[2] For instance, immigration is very often blamed for increasing unemployment, diverting public spending away from citizens, and provoking a cultural war. These narratives, however, seem to overlook many important and relevant factors that explain unemployment, public spending or social identities. Is this misspecification the result of bounded rationality, incomplete information or can it have instrumental value?

In this paper, we argue that model misspecification can have *instrumental value* because it acts as a commitment device in strategic communication games between a Receiver and a Sender. We introduce a two-dimensional cheap talk game in which the Receiver faces two types of uncertainty. First, she does not

[1]This chapter is joint Agathe Pernoud. We are grateful to Francis Bloch, Matthew Gentzkow, Matthew Jackson, Philippe Jehiel, Frédéric Koessler, Paul Milgrom, Ronny Razin, Olivier Tercieux for helpful conversations and comments. We also thank seminar participants at PSE, Stanford, Akbarpour–Milgrom discussion group for valuable comments and questions. S. Gleyze acknowledges the support of the EUR grant ANR-17-EURE-0001.

[2]The idea that populism is characterized by the simplicity of its narratives received empirical support in Bischof and Senninger (2018) who studies political manifestos in Austria and Germany between 1945 and 2013.

know which variables are payoff-relevant—this is referred to as "model uncertainty". Second, Receiver does not know the realization of these variables—which is the more standard "state uncertainty". Receiver communicates with a Sender who is informed about the state. To fix ideas, let Receiver be a politician (e.g., the President) and Sender be a media.[3] The politician would like to approve policies that address some economic problem of interest. The success of each policy depends on which variables she thinks are the cause of the problem (her "worldview")[4] and their realization (the realized state). The media only has partially aligned preferences with the politician. We first take as given Receiver's worldview and study how it affects communication with the Sender. We then endogenize her worldview by introducing another agent, the Principal (e.g. a voter), who observes the true model of the world and can either choose Receiver with a specific worldview, or communicate about the worldview beforehand.

Our first main result is that, in the communication game between Sender and Receiver, holding a misspecified worldview can actually increase informativeness of communication in equilibrium. The intuition is that simple models—i.e., believing that few variables are actually relevant—reduce the number of individually rational actions for Receiver, which can lead to more communication as there is less room for information manipulation in equilibrium. We then show how holding such a model, even if it is misspecified, can be welfare improving for Receiver.

Second, we endogenize Receiver's worldview by introducing a third party, a Principal, who is informed about the true model of the world. We show that if the Principal can delegate decision-making by choosing a Receiver with a specific worldview, then he will always choose a Receiver with a simpler, monocausal worldview. Hence model misspecification, defined as Receiver holding an incorrect worldview given the true realized model, can have instrumental value in strategic communication. Framed in the context of our running example, this suggests voters may elect populist politicians not because they agree

[3] In the paper we also consider an application in which Receiver is a CEO and Sender is a syndicate.

[4] We use the terms "worldview" and "model" interchangeably throughout the paper.

with their worldviews, but because they believe such candidates will not be "pushovers" that are easily influenced by lobbyists, the media or the administration.[5] Similarly, if the Principal can only communicate on models (so he cannot directly choose the worldview of Receiver, but can try to influence it via communication), we show that all equilibria are outcome equivalent to a babbling equilibrium—namely, communication on models is impossible. This is precisely driven by the fact that the Principal benefits from Receiver holding a misspecificed worldview, which prevents meaningful communication.

Related Literature We contribute to the literature on (multidimensional) cheap talk. Crawford and Sobel (1982) introduce the canonical model of strategic communication without commitment. Under the assumption of state independent preferences, Chakraborty and Harbaugh (2010) extend the basic model to a multi-dimensional setting and prove equilibrium existence. Lipnowski and Ravid (2020) further provide a geometric characterization of equilibrium payoffs. Levy and Razin (2007) show that the correlation structure between dimensions of the state puts bounds on equilibrium communication. Instead, we do comparative statics with respect to the set of individually rational actions, fixing the correlation structure. The main innovation of the present paper is to introduce subjective models—describing which variables Receiver should act on—and belief updating on such models. Of course, what we call "models" could be embedded in a larger state space, but we think that the conceptual distinction that we introduce is relevant and allow to study more specifically the formation of worldviews. Related to this literature on cheap talk and the present paper, Che and Kartik (2009) show that in a disclosure game the Receiver may choose a Sender with a different prior to incentivize information

[5]Most common explanations of populism borrow concepts from identity politics. (See Guriev and Papaioannou (2022) for an extensive review.) For instance, Eichengreen (2018) argues that populism rises in times of economic crisis because "elites", the winners of the preceding period, are unwilling or unable to redistribute with the "losers". Norris and Inglehart (2019) instead advocate for a theory of cultural backlash: populism rises because a majority group (e.g. white males) feel endangered by the empowerment of women, and the support for underprivileged ethnic, racial, and religious groups—which are seen as a threat to their identity. To the best of our knowledge, there is no investigation of the instrumental value of populism to date.

acquisition.

Schwartzstein and Sunderam (2021) introduce the idea of communication on models by having Receiver change her worldview whenever past data are more likely under the competing model than under the default model. They show that there is a trade-off between how well the new model fits past data, and movement in beliefs—namely, when data are unsurprising under the competing model, the agent does not update her belief. In their model, Receiver has no prior on subjective causal models and communication is not strategic, whereas in our model Receiver is Bayesian and communication is strategic. This leads to different equilibrium predictions, as in our setting we show that communication on models is impossible, whereas in Schwartzstein and Sunderam (2021) it is easy to change Receiver's worldview even when the default model is correct.

Eliaz and Spiegler (2020) show that when agents select causal models by maximizing anticipatory utility, they tend to choose misspecified models. Olea et al. (2022) show that agents that have simple models, i.e. use few variables to predict an outcome, have more confidence in their estimate when sample size is small. Levy et al. (2022) introduce a model of political competition between two groups, where one group has a simpler subjective model than the other one. They show that this leads to policy cycles and extreme policy choices. The main difference with these papers and ours is that we consider the impact of misspecified models on strategic communication.

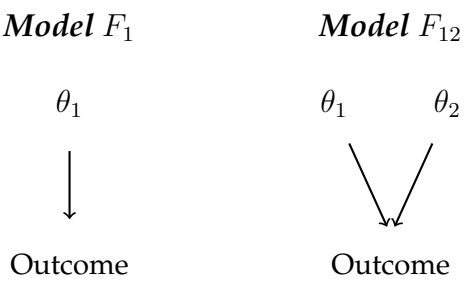
Finally, observe that our concept of worldview can be interpreted as a commitment device on individually rational actions. Therefore, our paper relates to the literature studying Receiver's commitment power. Most notably, Dessein (2002) shows that a principal prefers to delegate decision making to an informed Sender whenever the conflict of interest is small—which can be interpreted as a commitment to an action rule conditional on a message. The main difference with our paper is that a worldview is a belief that can evolve, hence we can investigate novel questions such as communication on models. We further discuss the relation between our model and one in which Receiver has commitment power in Section 4.

4.1 How Models Shape Communication

4.1.1 Setup

We consider a cheap talk game with two types of uncertainty. Receiver faces "model uncertainty": she does not know which variables, i.e., which dimensions of the state space, are payoff-relevant. In reality, she probably has uncertainty on the entire joint distribution of variables, but as a first step we only consider uncertainty on *which* variables she should care about. Moreover, Receiver also faces the traditional "state uncertainty", i.e. uncertainty about the realization of these variables. In this section, we consider the case of a unique Sender, who is informed about the state and can communicate on it.

Let $\Theta = \{\underline{\theta}_1, \bar{\theta}_1\} \times \{\underline{\theta}_2, \bar{\theta}_2\} \equiv \{0,1\}^2$ be the state space, so that the state is composed of two variables—θ_1 and θ_2—each of which can either be high or low. A *model* specifies which of these variables cause an economic outcome of interest. (We use the terms (causal) model and worldview interchangeably throughout the paper.) There are two possible models:[6]



According to Model F_1, only variable θ_1 is causing the outcome of interest, whereas under Model F_{12} both variables are causing the outcome. The true model of the world is unknown to Receiver. Receiver believes that the true model is F_1 with probability λ, and is F_{12} with complementary probability $1-\lambda$. This section investigates how Receiver's worldview, i.e., the relative weight λ she puts on Model F_1, affects communication in equilibrium.

[6]Causal models can be thought of as directed acyclic graphs (DAGs) in which a variable is an ancestor of another in the graph if it is one of its cause.

Receiver can take action to address each of the two possible causes. Let $A = \{\underline{a}_1, \overline{a}_1\} \times \{\underline{a}_2, \overline{a}_2\} \equiv \{0,1\}^2$ be the action space. Interpret the high action $\overline{a}_k$ as Receiver taking active measures to reduce variable θ_k, and the low action $\underline{a}_k$ as the status quo. Receiver only wants to act on the true causes of the problem. If the true model is F_1, the optimal action on the first dimension is $a_1^*(\theta_1, F_1) = \overline{a}_1$ if $\theta_1 = \overline{\theta}_1$ and $a_1^*(\theta_1, F_1) = \underline{a}_1$ if $\theta_1 = \underline{\theta}_1$. It is however optimal to set $a_2^*(\theta_2, F_1) = \underline{a}_2$ irrespective of θ_2 as that variable does not contribute to unemployment under model F_1. It is *as if* high actions were costly—hence if a variable is irrelevant then the status-quo is optimal. On the contrary, if the true model is F_{12}, then the optimal action along both dimensions is to match the state.

Instead of defining Receiver's preferences on final outcomes, we directly define preferences on actions, states and models. We show in Section 4 how these reduced-form preferences can be microfounded using final outcomes. Receiver's payoff from actions (a_1, a_2) in state (θ_1, θ_2) if the true model is F writes:[7]

$$u_R(a_1, a_2, \theta_1, \theta_2, F) = -(a_1 - a_1^*(\theta_1, F))^2 - (a_2 - a_2^*(\theta_2, F))^2.$$

Sender's preferences are only partially aligned with Receiver's:

$$u_S(a_1, a_2, \theta_1, \theta_2, F) = -(a_1 - a_1^*(\theta_1, F))^2 + \gamma a_2$$

with $\gamma > 1$. Namely, Sender and Receiver are aligned on the first issue θ_1 but Sender has an agenda on the second issue θ_2 and wants higher action regardless of the true model or the realized state. This misalignment might prevent full communication about the state in equilibrium, and is key for our analysis.

The joint distribution of states is

with $\mu_0 < 0.5$. For tractability purposes, we restrict attention to the case of perfect correlation between the two variables. We discuss at the end of the paper how our results generalize to settings with negative correlation.

[7] Assuming quadratic preferences for Receiver simplifies the exposition greatly, but it does not seem to be driving our results. Most of them extend when the gains from taking the correct action depend on the dimension, the state, and the model.

	$\theta_2 = 0$	$\theta_2 = 1$
$\theta_1 = 0$	$1 - \mu_0$	0
$\theta_1 = 1$	0	μ_0

The timing of the game is as follows: First, Sender observes the state and sends a message $m \in M$ to Receiver. Sending messages is free, and Sender cannot commit to a specific communication protocol ex ante—hence this is a cheap-talk game. Receiver then takes an action. A (perfect Bayesian) equilibrium consists of a strategy $q_S : \Theta \longrightarrow \Delta M$ for Sender, a strategy $p_R : M \longrightarrow \Delta A$ for Receiver, and a belief system such that (i) Receiver's beliefs are derived from the prior μ_0 and q_S using Bayes' rule whenever possible, (ii) Receiver only plays actions that are optimal given her belief, and (iii) Sender only sends messages that maximize his expected utility given θ. We take the belief that Sender and Receiver assign to models λ as given. We focus on the Sender-preferred equilibrium.

Example 1 (Political Economy): Receiver is an elected politician who decides on policies (e.g., the President). Sender is a media who communicates on the state of the economy. The outcome of interest is unemployment, which has two possible causes: the extent of immigration θ_1 and of aggregate consumption θ_2. The politician aims at reducing unemployment, and to that end wants to address whichever variable(s) cause(s) it. To address immigration, she can impose additional legal requirements for new entrants in the country or limit more aggressively illegal immigration ($a_1 = 1$). To address low aggregate consumption, she can undertake an expansionary fiscal policy ($a_2 = 1$). The media wants greater public spending regardless of whether it is sufficiently high already, and whether low aggregate consumption is actually contributing to unemployment. Because of this agenda, the media tries to influence policy by exaggerating how low aggregate demand is.

Example 2 (Organization): Receiver is an employer who manages a firm (e.g., the CEO). Sender is an employee, or a worker syndicate, who communicates on issues faced by workers in the production process. The outcome of interest is (low) productivity, which has two possible causes: skill mismatch between

workers and the task θ_1 and workers' effort θ_2. The CEO aims at increasing the firm's productivity, and to that end wants to address whichever variable(s) depress(es) it. To address skill mismatch, she can provide additional training to the workers ($a_1 = 1$). To incentivize higher effort, she can increase wages ($a_2 = 1$). The worker syndicate wants higher wages regardless of whether they are sufficiently high already, or whether they are at all related to productivity. Because of this agenda, the syndicate tries to influence the CEO's decisions by exaggerating how costly effort is for workers.

4.1.2 Equilibrium Characterization

Let μ be the posterior probability that Receiver assigns to state $\theta = (1, 1)$. Receiver's optimal action as a function of μ and λ is

$$\sigma^*(\mu, \lambda) = \begin{cases} (0,0) & \text{if } \mu \leq \frac{1}{2} \\ (1,0) & \text{if } \mu \geq \frac{1}{2} \text{ and } \mu \leq \mu^* \equiv \frac{1}{2(1-\lambda)} \\ (1,1) & \text{if } \mu \geq \mu^* \end{cases}$$

Because the first variable (θ_1) is causing the outcome under both model F_1 and F_{12}, whether or not Receiver wants to act on it only depends on her belief about the state: She sets $a_1 = 1$ if and only if she thinks that the high state is sufficiently likely, i.e., $\mu \geq 0.5$. On the contrary, the second variable (θ_2) only contributes to unemployment under model F_{12}. Hence Receiver wants to take a high action $a_2 = 1$ only if she puts sufficient weight on both the high state *and* model F_{12}.

First, note that for $\lambda > 0.5 \equiv \lambda^*$, Receiver never takes action $a_2 = 1$. Hence the only actions that Sender can induce are $a = (0, 0)$ and $a = (1, 0)$, and since his preferences are aligned with Receiver's over those, his most preferred equilibrium is fully revealing.

Second, for $\lambda \leq \lambda^*$ it is possible to induce action $a_2 = 1$. What is not possible, however, is that in equilibrium Sender sends a message m that induces actions $a = (1, 1)$ with probability one. Sending that message would ensure Sender a

payoff of

$$-1 + \gamma > 0$$

in state $\theta = (0, 0)$, and Sender would *always* want to send it.

The only informative equilibrium then has Sender send two messages m_1 and m_0. Message m_0 indicates the low state with certainty, and $\Pr(a = (0, 0) \mid m_0) = 1$. Message m_1 leads Receiver to randomize between actions $(1, 1)$ and $(1, 0)$. In state $\theta = (0, 0)$, Sender must then be indifferent between inducing action $(0, 0)$ for sure and inducing a lottery over actions $(1, 1)$ and $(1, 0)$, which requires

$$\begin{aligned} &\Pr(a = (1,1) \mid m_1)\mathbb{E}_\lambda[u_S(\overline{a}_1, \overline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] \\ &\quad + \Pr(a = (1,0) \mid m_1)\mathbb{E}_\lambda[u_S(\overline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] = \mathbb{E}_\lambda[u_S(\underline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] \\ \iff &\Pr(a = (1,1) \mid m_1)[-1 + \gamma] - \Pr(a = (1,0) \mid m_1) = 0. \end{aligned}$$

Namely,

$$\Pr(a = (1, 1) \mid m_1) = \frac{1}{\gamma}.$$

Since Receiver must randomize between actions $(1, 1)$ and $(1, 0)$ upon receiving m_1, this requires her to be indifferent between taking the low and high action a_2: her belief about the high state must be precisely equal to μ^*. To close the equilibrium, we now derive the strategy of Sender that induces this belief upon observing m_1:

$$\Pr(\theta = (1,1) \mid m_1) = \frac{\Pr(m_1 \mid \theta = (1,1))\mu_0}{\Pr(m_1 \mid \theta = (1,1))\mu_0 + \Pr(m_1 \mid \theta = (0,0))(1 - \mu_0)} = \mu^*.$$

Overall, the communication strategy of Sender is

$$q_S(m_1 \mid \theta = (1,1)) = 1, \qquad q_S(m_1 \mid \theta = (0,0)) = \frac{\mu_0}{1 - \mu_0}\frac{1 - \mu^*}{\mu^*}$$

and m_0 with complementary probability. In response, Receiver chooses the

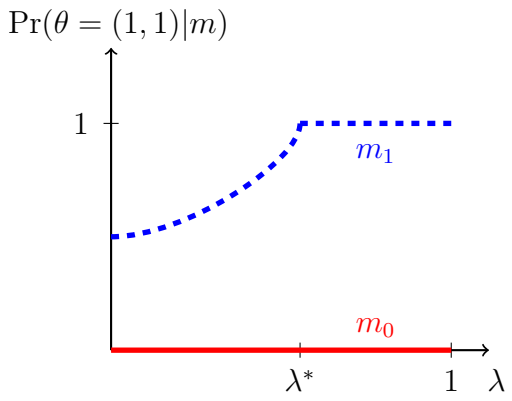


Figure 4.1: *Posterior beliefs of Receiver in equilibrium conditional on m_0 and m_1 as a function of $\lambda = \Pr(F_1)$.*

following distribution over actions

$$p_R(a = (0,0) \mid m_0) = 1, \quad p_R(a = (1,1) \mid m_1) = \frac{1}{\gamma}, \quad p_R(a = (1,0) \mid m_1) = 1 - \frac{1}{\gamma}.$$

Our first main result is that holding a simple worldview (i.e., putting more weight on the single-cause model F_1) makes the equilibrium more informative, as can be seen from Figure 4.1.

PROPOSITION 19. *The informativeness (in the Blackwell sense) of the equilibrium is monotonically increasing in λ.*

This follows directly from the above characterization of equilibrium behavior: The probability that Sender sends the "wrong" message m_1 in state $\theta = (0,0)$ decreases in λ. Indeed, if Receiver puts a lot of weight on the simple model F_1, then she needs to be almost certain that the state is high to be willing to take action $a_2 = \overline{a}_2$. Hence more communication is required.

Next, we show that Receiver benefits from holding the simple worldview F_1. Let $V(\lambda)$ denote the expected utility of Receiver in equilibrium, as a function of the probability she assigns to model F_1:

$$V(\lambda) \equiv \sum_{\theta} \mu_0(\theta) \sum_{m} q_S(m|\theta;\lambda) \sum_{a} p_R(a|m;\lambda) \mathbb{E}_\lambda[u_R(a,\theta,F)].$$

Note, however, that this is not a straightforward implication of Blackwell's

theorem as λ changes both equilibrium communication *and* Receiver's preferences. Indeed, λ changes how much Receiver weights her expected utility conditional on model F_1 against model F_{12}. The next result shows that, overall, putting more weight on the simple model F_1 is welfare improving for Receiver.

PROPOSITION 20. *The expected utility of Receiver $V(\lambda)$ is monotonically increasing in λ.*

The worldview that Receiver holds λ has two effects: it impacts equilibrium play—q_S and p_R,—as well as how Receiver evaluates the outcome induced by equilibrium play—$\mathbb{E}_\lambda[u_R(\cdot)]$. As λ increases, the equilibrium becomes more informative (Proposition 1) and allows Receiver to better target action a_1 to the realized state θ_1. The same is however not true for action a_2, as when λ goes above λ^* Receiver stops taking action $a_2 = \overline{a}_2$ altogether, which is costly if the true model is F_{12}. Hence Receiver trades-off better decision-making on the first dimension, with potentially more mistakes on the second. Since a greater λ also means that Receiver puts more weight on model F_1 when evaluating the equilibrium outcome, not taking action $a_2 = \overline{a}_2$ is less likely to be a mistake, and her overall expected payoff is larger.

4.2 The Value of Model Misspecification

In the previous section, we showed that Receiver's expected utility is increasing in her subjective belief in the simple model F_1. This is due to two effects: a strategic effect (informativeness of communication increases), and a preference effect (Receiver believes she is doing fewer mistakes). In this section, we disentangle them, and show that even if we neutralize the preference effect by looking through the lens of an informed Principal, the Principal is still better off with Receiver holding a simpler misspecified model. Therefore, this shows that misspecification can have a positive value in strategic communication.

4.2.1 Delegation

For the following two sections, we extend our framework and introduce another agent, a Principal. The Principal is informed of the true model F, but

does not know the state θ. Instead of communicating directly with Sender and making decisions himself, the Principal can delegate decision-making to an Agent who is the Receiver from the previous section. The latter then communicates on states with Sender in a second stage. Receiver and the Principal share the same preferences, but not the same worldview: Principal knows the true model whereas Receiver puts weight λ on model F_1.

Let $V_P(\lambda|F)$ denote the expected equilibrium payoff of Principal when the true model is F and he delegates decision-making to a Receiver with worldview λ. Using the above characterization of equilibrium communication between Sender and Receiver, these write

$$V_P(\lambda|F_{12}) \equiv \sum_\theta \mu_0(\theta) \sum_m q_S(m|\theta;\lambda) \sum_a p_R(a|m;\lambda) u_R(a,\theta,F_{12})$$
$$V_P(\lambda|F_1) \equiv \sum_\theta \mu_0(\theta) \sum_m q_S(m|\theta;\lambda) \sum_a p_R(a|m;\lambda) u_R(a,\theta,F_1)$$

Figure 4.2 plots these expected payoffs.

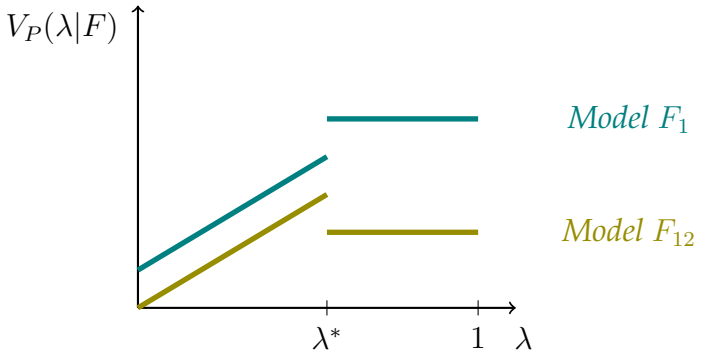


Figure 4.2: *Principal's equilibrium expected payoff as a function of Receiver's belief on models λ, for each possible realization of the true model.*

We assume Principal can choose the worldview of Receiver—perhaps because there is a vast pool of agents with various worldviews from which Principal can hire. Say Receiver's model is misspecificed if her belief about models is incorrect, that is if $\lambda > 0$ while the true model is F_{12}, or if $\lambda < 1$ while the true model is F_1.

PROPOSITION 21. *When the true Model is F_1, the Principal chooses a Receiver with $\lambda = 1$. When the true Model is F_{12}, the Principal optimally chooses a Receiver with a misspecified, simpler model:*

$$\arg\max_{\lambda} V_P(\lambda|F_{12}) = \lambda^* > 0.$$

Therefore, even when the Principal knows that the true model is complex and multi-causal F_{12}, he chooses a Sender who puts significant weight on the simpler, uni-causal model. Indeed, holding such misspecified worldview yields Receiver a higher expected payoff when communicating with Sender, as it attenuates how much Sender tries to mislead Receiver in equilibrium to serve his own agenda. This is true even though a misspecified worldview leads Receiver to take actions that are ex post suboptimal in the eyes of the Principal, as it makes Receiver "too conservative" when it comes to taking action $a_2 = \overline{a}_2$.

In example 1, the Principal can be thought of as a voter who chooses a populist representative in the hope that he will not get fooled by the media or lobbyists. Proposition 3 says that voting for such populist representative is optimal even when the voter does not share the populist's worldview, but instead believes in a more complex, multi-causal model of the world. In example 2, the Principal can be thought of as a board of directors who chooses a conservative CEO who does not believe low wages can ever contribute to low productivity, and will not be easily manipulated by syndicates. Again, choosing such a CEO is optimal even when the board of directors does not share the CEO's worldview.

4.2.2 Communication on Models

Now suppose that Principal is not able to directly *choose* the worldview of Receiver—perhaps the pool of agents is not that rich, or the agent's worldview is hard to identify at the time of the hire. What Principal can however do is communicate about the true model with Receiver. For instance, in Example 2, the board of directors can share with the CEO their accumulated knowledge of the workings of the firm, and in particular of how much effort and efficient wage levels affect its overall productivity.

The timing of the game is as follows. First, the true model is drawn according to a prior distribution $\lambda_0 = \Pr(F_1)$ and is observed by the Principal. Principal sends a message $m \in M$ from some arbitrary set of messages. To keep the analysis simple and uncluttered, suppose that this message is public, in the sense that it is observed by both Receiver and State-Sender. In a second stage, the state is drawn and observed by State-Sender. The game then unfolds as before. The previous analysis hence characterizes what happens in this second stage of communication, given some belief λ that resulted from communication with Principal. We solve for the equilibrium communication on models in the first stage.

As before, the expected payoff of Principal when Receiver holds posterior λ while the realized model is F is $V_P(\lambda|F)$. Let $q_P : \{F_1, F_{12}\} \longrightarrow \Delta M$ denote Principal's communication strategy. Receiver's belief about models upon receiving m is derived from Bayes' law:

$$\lambda(m) = \frac{q_P(m|F_1)\lambda_0}{q_P(m|F_1)\lambda_0 + q_P(m|F_{12})(1-\lambda_0)}$$

for all $m \in \operatorname{supp} q_P \equiv \{m|q_P(m|F) > 0 \text{ for some } F\}$. For q_P to be an equilibrium, we must have that for all F, $q_P(\cdot|F)$ is supported on

$$\arg\max_{m \in M} V_P(\lambda(m)|F).$$

If not, Principal must sometimes be sending a strictly suboptimal message for some realization of the model F, and must hence have an incentive to deviate. This in particular implies that if Principal sends several messages with positive probability under model F, then he must be indifferent between sending all such messages: $V_P(\lambda(m)|F) = V_P(\lambda(m')|F)$ for all $m, m' \in \operatorname{supp} q_P(\cdot|F)$.

Can Principal provide meaningful information about the realized model to Receiver? We show that the answer is no: Principal benefits from Receiver holding a misspecified model, which prevents credible communication.

PROPOSITION 22. *In any equilibrium, communication on models in the first stage is payoff-equivalent to a babbling equilibrium.*

Figure 4.2 is helpful to understand why communication is impossible. Any

informative equilibrium must involve at least one message m_{12} that indicates model F_{12} is more likely—namely, $\lambda(m_{12}) < \lambda_0$—and another m_1 that indicates the opposite—$\lambda(m_1) > \lambda_0$. If the true model is F_1, Principal has a strict incentive to be truthful as soon as $\lambda(m_{12}) \leq \lambda^*$.[8] But then Receiver knows that message m_{12} can never be sent when $F = F_1$, and so his posterior worldview must be $\lambda(m_{12}) = 0$. This is however the payoff-*minimizing* worldview under both realizations of the model, as Receiver benefits from putting some weight on model F_1 even when the true model is F_{12}. Hence, upon observing F_{12}, Principal has a strict incentive to deviate and send whichever message indicates F_1 is more likely. Despite the facts that Principal's preferences are fully aligned with Receiver's, communication is very limited so as to prevent Receiver from being manipulated in the second stage by State-Sender.

4.3 Discussion

Misspecified Models vs. Commitment Power. Holding a misspecified model of the world can be valuable in strategic communication as it allows Receiver to be more conservative when it comes to taking actions towards which Sender is biased. One can think of this as a form of commitment power: a Receiver who puts more weight on model F_1 needs more evidence that the high state $\theta = (1, 1)$ has realized to be willing to take action $a_2 = \overline{a}_2$. Of course, worldviews give Receiver much less flexibility than if she could fully commit to a decision rule $\rho : M \longrightarrow \Delta A$, as Receiver's actions must still be consistent with her posterior about the state. Yet, we show that, in our setting, the optimal decision rule under full commitment can be implemented by holding a specific worldview, and behaving optimally according to that worldview.

PROPOSITION 23. *Suppose the true model is F_{12}.*[9] *When she holds the optimal mis-*

[8] If $\lambda(m_{12}) > \lambda^*$, i.e., if Receiver's posterior always puts a weight greater than λ^* on model F_1, then equilibrium communication with State-Sender is unaffected by first-stage communication. Indeed, for all $\lambda > \lambda^*$, State-Sender fully reveals the state in equilibrium and Receiver never takes action $a_2 = 1$.

[9] When the true model is F_1, commitment power has no value. Indeed, Receiver then never acts on the second issue and always sets $a_2 = 0$. Since Sender and Receiver share the same preferences over the first issue, the equilibrium is fully revealing and cannot be improved upon

specified worldview $\lambda = \lambda^*$, Receiver achieves the same expected equilibrium payoff as if she had full commitment power.

If Receiver could fully commit to a decision rule, then without loss she would incentivize Sender to fully reveal the state, and would take the optimal action as often as possible while respecting this incentive compatibility constraint. To that end, she would commit to mix between $a = (1, 0)$ and $a = (1, 1)$ when Sender tells her the state is high, to ensure Sender remains truthful when the state is low. This coincides precisely with the equilibrium derived above when Receiver puts weight $\lambda = \lambda^*$ on model F_1. Hence one can think of a worldview as one way of implementing the optimal decision rule.

Negative Correlation and Independence Across States. All our analysis extends to the case of perfect negative correlation, i.e. $\Pr(\theta = (0, 1)) = \mu_0$ and $\Pr(\theta = (1, 0)) = 1 - \mu_0$. Whenever communication is feasible, SS's preferred equilibrium has again two messages m_0 and m_1 with

$$\Pr(m_1 \mid \theta = (0, 1)) = 1 \qquad \Pr(m_1 \mid \theta = (1, 0)) = \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*}$$

and m_0 with the complementary probability. In response, Receiver chooses the following distribution over actions

$$\Pr(a = (1, 0) \mid m_0) = 1 \quad \Pr(a = (0, 1) \mid m_1) = \frac{1}{\gamma} \quad \Pr(a = (0, 0) \mid m_1) = 1 - \frac{1}{\gamma}.$$

More generally, the analysis extends to any setting with sufficiently high correlation across states. This is due to the fact that when correlation is high, the equilibrium is constrained to be a monotonic partition, i.e. there is a message that indicates low states and another that indicates high states. Hence, increasing λ makes the incentive constraint tighter and necessarily leads to more information revelation. When variables are independent, however, there exist equilibria that do not have this monotonic structure. Hence, increasing λ do not necessarily lead to more information revelation. This means that while our result is a proof of principle that ignorance of the true model can have value

by commitment power.

in strategic communication, this needs not extend to all environments. There must be something linking the two types of actions together—e.g., correlation between θ_1 and θ_2—for the intuition behind our result to go through.

Micro-Foundation of Preferences. Let $y \in Y \subseteq \mathbb{R}$ denote the outcome variable of interest, e.g. the gap between the natural rate of unemployment and the current rate of unemployment. Let $c_k > 0$ denote the cost associated with taking action $a_k = \overline{a}_k$. A model specifies a data generating process, that is how the distribution of y depends on the realized state and the action taken by Receiver. Formally, there exist model-dependent probability measures $\Pr(\cdot \mid F)$ over $Y \times \Theta \times A$. If a variable does not belong to the model $k \notin F$, then it is as if the variable θ_k were not causally related to y, such that the distribution of y is independent of dimension k: $\Pr(y \mid \theta, a; F) = \Pr(y \mid \theta_{-k}, a_{-k}; F)$. If a variable belongs to the model $k \in F$, then it is a cause of unemployment and impacts the distribution of y.

Receiver has two possible actions associated with each variable k: a low action $\underline{a}_k$, which is costless; and a high action $\overline{a}_k$, which costs c_k. Taking the low action should be interpreted as maintaining the "status-quo," or equivalently remaining passive, as opposed to actively addressing variable k, which requires time and effort and is hence costlier. When all relevant variables are low—i.e., $\theta_k = \underline{\theta}_k$ for all $k \in F$—unemployment is as low as possible, and the gap between natural unemployment and current unemployment is zero. In this case, actions have no impact: $\mathbb{E}[y \mid \underline{\theta}, a; F] = 0$ for all $a \in A$. A high realization of a payoff-relevant variable $k \in F$ induces the following expected outcome: $\mathbb{E}[y \mid \overline{\theta}, \underline{a}; F] < 0 = \mathbb{E}[y \mid \overline{\theta}, \overline{a}; F]$. Overall, the gap between natural unemployment and current unemployment can be written as

$$\mathbb{E}[y \mid \theta, a; F] = -\sum_{k \in F} \mathbb{1}\{\theta_k = \overline{\theta}_k, a_k = \underline{a}_k\}.$$

Receiver's payoff equals this gap net of action costs:

$$u_R(a, \theta, F) = \mathbb{E}[y \mid \theta, a; F] - \sum_{k=1,2} c_k \mathbb{1}\{a_k = \overline{a}_k\}.$$

The reduced-form preferences that we use in the main analysis are equivalent to the above preferences with $c_k = 0.5$ for $k = 1, 2$.

Proofs

Proof of Proposition 20. Receiver's expected payoff in equilibrium equals

$$\begin{aligned}
V(\lambda) = \mathbf{1}\{\lambda \leq \lambda^*\} \Bigg[& \mu_0 \Big(\frac{1}{\gamma} \mathbb{E}_\lambda[u(\overline{a}_1, \overline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F)] + \frac{\gamma - 1}{\gamma} \mathbb{E}_\lambda[u(\overline{a}_1, \underline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F)] \Big) \\
& + (1 - \mu_0) \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*} \Big(\frac{1}{\gamma} \mathbb{E}_\lambda[u(\overline{a}_1, \overline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] + \frac{\gamma - 1}{\gamma} \mathbb{E}_\lambda[u(\overline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] \Big) \\
& + (1 - \mu_0) \left(1 - \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*} \right) \mathbb{E}_\lambda[u(\underline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] \Bigg] \\
& + \mathbf{1}\{\lambda > \lambda^*\} \Big[\mu_0 \mathbb{E}_\lambda[u(\overline{a}_1, \underline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F)] + (1 - \mu_0) \mathbb{E}_\lambda[u(\underline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F)] \Big] \\
= \mathbf{1}\{\lambda \leq \lambda^*\} & \Big[- \mu_0 \Big(\frac{\lambda}{\gamma} + \frac{(\gamma - 1)(1 - \lambda)}{\gamma} \Big) - \mu_0 \frac{1 - \mu^*}{\mu^*} \Big(\frac{2}{\gamma} + \frac{\gamma - 1}{\gamma} \Big) \Big] \\
+ \mathbf{1}\{\lambda > \lambda^*\} & \Big[- \mu_0 (1 - \lambda) \Big].
\end{aligned}$$

Using the fact that $\mu^* = \frac{1}{2(1-\lambda)}$, this simplifies to

$$V(\lambda) = -\mathbf{1}\{\lambda \leq \lambda^*\} \mu_0 (2 - 3\lambda) - \mathbf{1}\{\lambda > \lambda^*\} \mu_0 (1 - \lambda).$$

It is then easily verified that $V(\lambda)$ is monotonically increasing over $[0, \lambda^*]$ and $(\lambda^*, 1]$, and that it is continuous at $\lambda^* = 0.5$.

□

Proof of Proposition 21. We derive the expected payoff of Receiver in equilibrium given λ, for each possible realization of the true model:

$$\begin{aligned}
V(\lambda | F_{12}) = \mathbf{1}\{\lambda \leq \lambda^*\} \Bigg[& \mu_0 \Big(\frac{1}{\gamma} u(\overline{a}_1, \overline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F_{12}) + \frac{\gamma - 1}{\gamma} u(\overline{a}_1, \underline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F_{12}) \Big) \\
& + (1 - \mu_0) \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*} \Big(\frac{1}{\gamma} u(\overline{a}_1, \overline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F_{12}) + \frac{\gamma - 1}{\gamma} u(\overline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F_{12}) \Big) \\
& + (1 - \mu_0) \left(1 - \frac{\mu_0}{1 - \mu_0} \frac{1 - \mu^*}{\mu^*} \right) u(\underline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F_{12}) \Bigg]
\end{aligned}$$

$$+\mathbf{1}\{\lambda > \lambda^*\}\left[\mu_0 u(\overline{a}_1, \underline{a}_2, \overline{\theta}_1, \overline{\theta}_2, F_{12}) + (1-\mu_0)u(\underline{a}_1, \underline{a}_2, \underline{\theta}_1, \underline{\theta}_2, F_{12})\right]$$
$$= -\mathbf{1}\{\lambda \leq \lambda^*\}\frac{2\mu_0}{\gamma}[\gamma - (1+\gamma)\lambda] - \mathbf{1}\{\lambda > \lambda^*\}\mu_0.$$

$$V(\lambda|F_1) = -\mathbf{1}\{\lambda \leq \lambda^*\}\frac{\mu_0}{\gamma}[1 + (1-2\lambda)(\gamma+1)].$$

$V(\lambda|F_{12})$ is monotonically increasing in λ over $[0, 0.5]$, at which point it jumps downward to $-\mu_0$ and remains constant. □

Proof of Proposition 22. We proceed by backward induction, and first analyze what happens in stage 2—i.e., when Receiver communicates with State-Sender—given her posterior belief λ about model. For this we mostly rely on the analysis of Section 2. The equilibrium derived in Section 2 is the only informative one. All other equilibria are babbling and independent of Receiver's worldview: At the prior $\mu_0 < 0.5$, Receiver always takes action $a = (0, 0)$ irrespective of λ. Hence, if agents expected such babbling equilibrium to arise in stage 2, communication about models would be pointless, and hence payoff-equivalent to babbling.

From now on, let us then focus on the more interesting case in which agents anticipate that the Sender-preferred (and hence informative) equilibrium will be played in stage 2. Principal's expected payoff when the true model is F and Receiver has belief λ about models is then $V_P(\lambda|F)$. We show that all equilibria in the first stage are payoff-equivalent to a babbling equilibrium, in that they yield a payoff of $V_P(\lambda_0|F)$ to Principal.

In any informative equilibrium, Principal must send at least two messages with positive probability that lead Receiver to update her belief about models. More precisely, there must exist one message m_{12} that is sent with positive probability and that leads Receiver to update her belief downwards—$\lambda(m_{12}) < \lambda_0$—and another m_1 that leads Receiver to update her belief upward—$\lambda(m_1) > \lambda_0$. First consider the simpler case in which Receiver's posterior belief is always above λ^*: $\lambda(m) > \lambda^*$ for all $m \in \operatorname{supp} q_P$. Then, irrespective of what message P sends in the first stage, communication with SS in the second stage always yields the same outcome: SS fully reveals the state, and R sets $a_1 = \theta_1$ but $a_2 = 0$, always. The exact same outcome would have been achieved had P

not communicated any information about the model. Indeed, Receiver would have remained at her prior λ_0, which must be above λ^*,[10] and hence interacted in the exact same way with SS.

Now consider the more interesting case in which Receiver's posterior belief is sometimes below λ^*: there exist $m \in \operatorname{supp} q_{MS}$ such that $\lambda(m) \leq \lambda^*$ and call such message m_{12}. Note that sending such message when the true model is F_1 is strictly suboptimal as $V_P(\lambda|F_1) < V_P(\lambda'|F_1)$ whenever $\lambda \leq \lambda^*$ and $\lambda < \lambda'$. Hence Principal can never send message m_{12} when the true model is F_1 as sending whichever message m_1 leads Receiver to update her belief upward yields a strictly greater payoff: $q_P(m_{12}|F_1) = 0$. But then $\lambda(m_{12}) = 0$, that is, upon receiving message m_{12}, Receiver must know for sure that the true model is F_{12}. That however cannot occur in equilibrium as revealing fully that the model is F_{12} yields the lowest possible payoff for P: $V_P(0|F_{12}) < V_P(\lambda|F_{12})$ for all $\lambda > 0$. Hence P would want to deviate and send whichever message leads to a higher posterior.

□

Proof of Proposition 23. To derive the optimal decision rule under full commitment, we can rely on the Revelation Principle: It is without loss to restrict attention to equilibria under which the Sender announces a state $M = \Theta$ and reports truthfully $q_S(m = \theta|\theta) = 1$. The optimal decision rule $\rho : \Theta \longrightarrow \Delta A$ then solves

$$\begin{aligned}
\max_{\rho} & -(1-\mu_0)[2\rho(\overline{a}_1, \overline{a}_2|\underline{\theta}) + \rho(\overline{a}_1, \underline{a}_2|\underline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\underline{\theta})] \\
& \qquad - \mu_0[2\rho(\underline{a}_1, \underline{a}_2|\overline{\theta}) + \rho(\overline{a}_1, \underline{a}_2|\overline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\overline{\theta})] \\
\text{s.t.} \quad & -(\rho(\overline{a}_1, \overline{a}_2|\underline{\theta}) + \rho(\overline{a}_1, \underline{a}_2|\underline{\theta})) + \gamma(\rho(\overline{a}_1, \overline{a}_2|\underline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\underline{\theta})) \\
& \qquad \geq -(\rho(\overline{a}_1, \overline{a}_2|\overline{\theta}) + \rho(\overline{a}_1, \underline{a}_2|\overline{\theta})) + \gamma(\rho(\overline{a}_1, \overline{a}_2|\overline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\overline{\theta})) \\
\text{and} \quad & -(\rho(\underline{a}_1, \overline{a}_2|\overline{\theta}) + \rho(\underline{a}_1, \underline{a}_2|\overline{\theta})) + \gamma(\rho(\overline{a}_1, \overline{a}_2|\overline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\overline{\theta})) \\
& \qquad \geq -(\rho(\underline{a}_1, \overline{a}_2|\underline{\theta}) + \rho(\underline{a}_1, \underline{a}_2|\underline{\theta})) + \gamma(\rho(\overline{a}_1, \overline{a}_2|\underline{\theta}) + \rho(\underline{a}_1, \overline{a}_2|\underline{\theta}))
\end{aligned}$$

First note that under an optimal decision rule, the probability of taking the high action $a_2 = \overline{a}_2$ must be weakly higher in state $\theta = (1,1)$ than in state $\theta = (0,0)$,

[10] It is impossible to only induce posteriors $\lambda(m)$ that all lie strictly above the prior.

and so the only incentive compatibility constraint that binds is the first one.

Second, setting $\rho(\underline{a}_1, \overline{a}_2|\overline{\theta}) > 0$ cannot be optimal: any positive weight on action $a = (0,1)$ in the high state can be shifted to action $a = (1,1)$. This relaxes the incentive constraint and strictly increases the objective of Receiver. The same is true for for $\rho(\underline{a}_1, \underline{a}_2|\overline{\theta})$: any positive weight on action $a = (0,0)$ in the high state can be shifted to action $a = (1,0)$, yielding a strict improvement. A similar logic yields that, optimally, $\rho(\overline{a}_1, \overline{a}_2|\underline{\theta}) = 0$, $\rho(\overline{a}_1, \underline{a}_2|\underline{\theta}) = 0$, and $\rho(\underline{a}_1, \overline{a}_2|\underline{\theta}) = 0$. Hence, under an optimal decision rule, Receiver chooses action $a = (0,0)$ with probability one in state $(0,0)$.

The problem rewrites as

$$\max_{\rho} -\mu_0 \rho(\overline{a}_1, \underline{a}_2|\overline{\theta}) \quad \text{s.t. } 0 \geq -\rho(\overline{a}_1, \underline{a}_2|\overline{\theta}) + (\gamma - 1)\rho(\overline{a}_1, \overline{a}_2|\overline{\theta}).$$

Since $\rho(\overline{a}_1, \underline{a}_2|\overline{\theta}) + \rho(\overline{a}_1, \overline{a}_2|\overline{\theta}) = 1$ this yields

$$\rho(\overline{a}_1, \underline{a}_2|\overline{\theta}) = \frac{\gamma - 1}{\gamma}, \text{ and } \rho(\overline{a}_1, \overline{a}_2|\overline{\theta}) = \frac{1}{\gamma}.$$

The very same outcome is achieved without commitment, in the equilibrium derived in Section 2, for $\lambda = \lambda^*$. □

Bibliography

Abramowitz, A. I. and Saunders, K. L. (2008). Is Polarization a Myth? *The Journal of Politics*, 70(2):542–555.

Agarwal, N. and Somaini, P. (2018). Demand Analysis using Strategic Reports: An application to a School Choice Mechanism. *Econometrica*, 86(2):391–444.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alon, S. and Tienda, M. (2005). Assessing the "Mismatch" Hypothesis: Differences in College Graduation Rates by Institutional Selectivity. *Sociology of education*, 78(4):294–315.

Altmejd, A., Barrios Fernández, A., Drlje, M., Hurwitz, M., Kovac, D., Mulhern, C., Neilson, C., Smith, J., and Goodman, J. (2020). O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries.

Aumann, R. J. (1976). Agreeing to Disagree. *The Annals of Statistics*, 4(6):1236–1239.

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The quarterly journal of economics*, 107(3):797–817.

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics. *Journal of Public Economics*, 182:104123.

Becker, G. S. and Mulligan, C. B. (1997). The Endogenous Determination of Time Preference. *The Quarterly Journal of Economics*, 112(3):729–758.

Bergemann, D., Shi, X., and Välimäki, J. (2009). Information acquisition in interdependent value auctions. *Journal of the European Economic Association*, 7(1):61–89.

Bergemann, D. and Välimäki, J. (2002). Information acquisition and efficient mechanism design. *Econometrica*, 70(3):1007–1033.

Bernheim, B. D., Braghieri, L., Martínez-Marquina, A., and Zuckerman, D. (2021). A Theory of Chosen Preferences. *American Economic Review*, 111(2):720–54.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit Discrimination. *American Economic Review*, 95(2):94–98.

Bertrand, M. and Duflo, E. (2017). Field Experiments on Discrimination. *Handbook of Economic Field Experiments*, 1:309–393.

Bertrand, M., Hanna, R., and Mullainathan, S. (2010). Affirmative Action in Education: Evidence from Engineering College Admissions in India. *Journal of Public Economics*, 94(1-2):16–29.

Bettinger, E. P. and Long, B. T. (2005). Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. *American Economic Review P&P*, 95(2):152–157.

Bikhchandani, S. (2010). Information acquisition and full surplus extraction. *Journal of Economic Theory*, 145(6):2282–2308.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of political Economy*, 100(5):992–1026.

Bischof, D. and Senninger, R. (2018). Simple Politics for the People? Complexity in Campaign Messages and Political Knowledge. *European Journal of Political Research*, 57(2):473–495.

Bobkova, N. (2019). Information Choice in Auctions. Working paper.

Boissonnet, N. (2019). Rationalizing Preference Formation by Partial Deliberation. *PhD Thesis*.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2020). Cross-Country Trends in Affective Polarization. *NBER Working Paper*.

Caplin, A., Dean, M., and Leahy, J. (2017). Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy.

Chakraborty, A. and Harbaugh, R. (2010). Persuasion by Cheap Talk. *American Economic Review*, 100(5):2361–82.

Che, Y.-K. and Kartik, N. (2009). Opinions as Incentives. *Journal of Political Economy*, 117(5):815–860.

Cherepanov, V., Feddersen, T., and Sandroni, A. (2013). Rationalization. *Theoretical Economics*, 8(3):775–800.

Chung, K.-S. (2000). Role Models and Arguments for Affirmative Action. *American Economic Review*, 90(3):640–648.

Crawford, V. P. and Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, pages 1431–1451.

Crémer, J. and McLean, R. P. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pages 1247–1257.

Cripps, M. W. (2018). Divisible Updating.

De Clippel, G. and Eliaz, K. (2012). Reason-Based Choice: A Bargaining Rationale for the Attraction and Compromise Effects. *Theoretical Economics*, 7(1):125–162.

Dekel, E., Lipman, B. L., and Rustichini, A. (2009). Temptation-Driven Preferences. *The Review of Economic Studies*, 76(3):937–971.

Denti, T. (2018). Unrestricted Information Acquisition. Working Paper.

Dessein, W. (2002). Authority and Communication in Organizations. *The Review of Economic Studies*, 69(4):811–838.

Dietrich, F. and List, C. (2013a). A Reason-Based Theory of Rational Choice. *Nous*, 47(1):104–134.

Dietrich, F. and List, C. (2013b). Where Do Preferences Come From? *International Journal of Game Theory*, 42(3):613–637.

Dietrich, F. and List, C. (2016). Reason-Based Choice and Context-Dependence: An Explanatory Framework. *Economics & Philosophy*, 32(2):175–229.

Dominitz, J. and Manski, C. F. (1994). Eliciting Student Expectations of the Returns to Schooling. Technical report, National Bureau of Economic Research.

Eichengreen, B. (2018). *The Populist Temptation: Economic Grievance and Political Reaction in the Modern Era*. Oxford University Press.

Eliaz, K. and Spiegler, R. (2020). A Model of Competing Narratives. *American Economic Review*, 110(12):3786–3816.

Ellison, G. and Fudenberg, D. (1995). Word-of-Mouth Communication and Social Learning. *The Quarterly Journal of Economics*, 110(1):93–125.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*, volume 2. MIT press.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *American journal of sociology*, 91(3):481–510.

Gul, F. and Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, 69(6):1403–1435.

Gul, F. and Pesendorfer, W. (2005). The Revealed Preference Theory of Changing Tastes. *The Review of Economic Studies*, 72(2):429–448.

Guriev, S. and Papaioannou, E. (2022). The Political Economy of Populism. *Journal of Economic Literature*.

Hastings, J., Neilson, C. A., and Zimmerman, S. D. (2015). The Effects of Earnings Disclosure on College Enrollment Decisions. Technical report, National Bureau of Economic Research.

Hastings, J. S. and Weinstein, J. M. (2008). Information, School Choice, and Academic Achievement: Evidence from two Experiments. *The Quarterly journal of economics*, 123(4):1373–1414.

Hatfield, J. W., Kojima, F., and Kominers, S. D. (2018). Strategy-proofness, investment efficiency, and marginal returns: An equivalence. *Becker Friedman Institute for Research in Economics Working Paper*.

Hébert, B. M. and La'O, J. (2020). Information acquisition, efficiency, and nonfundamental volatility. Technical report, National Bureau of Economic Research.

Heller, Y. (2012). Justifiable choice. *Games and Economic Behavior*, 76(2):375–390.

Immorlica, N. S., Leshno, J. D., Lo, I. Y., and Lucier, B. J. (2018). Costly Information Acquisition and Stable Matching Mechanisms. Working Paper.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22:129–146.

Iyengar, S. and Westwood, S. J. (2015). Fear and Loathing Across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, 59(3):690–707.

Jehiel, P. (2005). Analogy-Based Expectation Equilibrium. *Journal of Economic theory*, 123(2):81–104.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, 125(2):515–548.

Kalai, G., Rubinstein, A., and Spiegler, R. (2002). Rationalizing Choice Functions by Multiple Rationales. *Econometrica*, 70(6):2481–2488.

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615.

Kapor, A. J., Neilson, C. A., and Zimmerman, S. D. (2020). Heterogeneous Beliefs and School Choice Mechanisms. *American Economic Review*, 110(5):1274–1315.

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2):132–157.

Larson, K. and Sandholm, T. (2001). Costly Valuation Computation in Auctions. *Theoretical Aspects of Rationality and Knowledge (TARK VIII)*, pages 169–182.

Levy, G. and Razin, R. (2007). On the Limits of Communication in Multidimensional Cheap Talk: A Comment. *Econometrica*, 75(3):885–893.

Levy, G., Razin, R., and Young, A. (2022). Misspecified Politics and the Recurrence of Populism. *American Economic Review*, 112(3):928–62.

Lipnowski, E. and Ravid, D. (2020). Cheap Talk with Transparent Motives. *Econometrica*, 88(4):1631–1660.

Manski, C. F. (1993). Adolescent Econometricians: How do Youth Infer the Returns to Schooling? In *Studies of Supply and Demand in Higher Education*, pages 43–60. University of Chicago Press.

Manski, C. F. (2004). Social Learning from Private Experiences: The Dynamics of the Selection Problem. *The Review of Economic Studies*, 71(2):443–458.

Mas-Colell, A. (1989). *The Theory of General Economic Equilibrium: A Differentiable Approach*. Cambridge University Press.

Matějka, F. and McKay, A. (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–98.

Mensch, J. (2019). Screening Inattentive Agents. Working Paper.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Nehring, K. (2006). Self-Control Through Second-Order Preferences. *Working Paper*.

Norris, P. and Inglehart, R. (2019). *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press.

Ok, E. A. (2007). *Real Analysis with Economic Applications*. Princeton University Press.

Olea, J. L. M., Ortoleva, P., Pai, M. M., and Prat, A. (2022). Competing models. *Quarterly Journal of Economics*.

Palacios-Huerta, I. and Santos, T. J. (2004). A Theory of Markets, Institutions, and Endogenous Preferences. *Journal of Public Economics*, 88(3-4):601–627.

Persico, N. (2000). Information Acquisition in Auctions. *Econometrica*, 68(1):135–148.

Ravid, D., Roesler, A.-K., and Szentes, B. (2019). Learning Before Trading: On the Inefficiency of Ignoring Free Information. Working Paper.

Ridout, S. (2021). Choosing for the right reasons. *Unpublished manuscript*.

Roesler, A.-K. and Szentes, B. (2017). Buyer-Optimal Learning and Monopoly Pricing. *American Economic Review*, 107(7):2072–80.

Rothstein, J. and Yoon, A. H. (2008). Affirmative Action in Law School Admissions: What do Racial Preferences Do? Technical report.

Schwartzstein, J. and Sunderam, A. (2021). Using Models to Persuade. *American Economic Review*, 111(1):276–323.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-Based Choice. *Cognition*, 49(1-2):11–36.

Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, 16(2):158–174.

Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3):665–690.

Strotz, R. H. (1955). Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies*, 23(3):165–180.

Tversky, A. and Simonson, I. (1993). Context-Dependent Preferences. *Management Science*, 39(10):1179–1189.

Zhong, W. and Bloedel, A. (2020). The Cost of Optimally Acquired Information. Working Paper.