Strategic Communication with Minimal Verification*

Gabriel Carroll[†] **Georgy Egorov**[‡]

April 2019

Abstract

A receiver wants to learn multidimensional information from a sender, and she has the capacity to verify just one dimension. The sender's payoff depends on the belief he induces, via an exogenously given monotone function. We show that by using a randomized verification strategy, the receiver can learn the sender's information fully in many cases. We characterize exactly when it is possible to do so. In particular, when the exogenous payoff function is submodular, we can explicitly describe a full-learning mechanism; when it is (strictly) supermodular, full learning is not possible. In leading cases where full learning is possible, it can be attained using an indirect mechanism in which the sender chooses the probability of verifying each dimension.

Keywords: Mechanism design, multidimensional information, verifiability, cheap talk

JEL Codes: D82, D83

*We are grateful to Sandeep Baliga, Vincent Crawford, Tommaso Denti, Xavier Gabaix, Johannes Hörner, Navin Kartik, Elliot Lipnowski, Paul Milgrom, Roger Myerson, Larry Samuelson, Ran Spiegler, Nisan Stiennon, and seminar participants at the University of Chicago, Northwestern University, the University of Pittsburgh, Princeton University, ETH/UZH, CEU, CERGE-EI, LMU and conference participants at Tel Aviv University, Cornell University, and Transatlantic Theory Workshop in Paris for helpful comments. Carroll also thanks the Research School of Economics at the Australian National University for their hospitality during a sabbatical. Authors are listed in random order.

[†] Stanford University. E-mail: gdc@stanford.edu

[‡] Kellogg School of Management, Northwestern University, and NBER. E-mail: g-egorov@kellogg.northwestern.edu

1 Introduction

An HR manager is interviewing a job candidate to form an opinion about the candidate's qualities or skills. A prosecutor is interviewing a defendant to decide whether there is a case that she could prosecute. An insurance company employee is evaluating a claim filed by its client to decide if it is legitimate or fraudulent. All these cases can be thought of as an interaction between a sender and a receiver of information, where the former tries to impress the latter, while the latter tries to infer the former's private information as precisely as possible.

This interaction is unlikely to be pure cheap talk. The HR manager can give the job candidate a test, or can call the college that the candidate lists on his vita to verify the truthfulness of the claim. The prosecutor can compare the defendant's statements with evidence obtained otherwise. The insurance company employee can visit the client's property and inspect what was damaged or stolen. However, the verification might be limited: the HR manager might be able to test only a few skills, the prosecutor might be able to corroborate only some of the defendant's claims, and the insurance company might verify only some of the information to ensure speedy processing.

In this paper, we make a strong assumption on the limits to verification: the sender's type is multidimensional, and the receiver is only able to verify one dimension. But the dimension she verifies can depend on the message that the sender sends. What can she do in this context?

The following example previews our ideas.

Example 1. An IT firm is hiring a programmer, and wants to evaluate a job candidate on two dimensions: math skills and coding skills. The candidate knows his skills x and y, but from the firm's perspective, they are i.i.d. uniform on $[0, 1]$. The candidate tries to impress the firm by signaling that the sum of his two skills, $x + y$, is as high as possible, because, for example, this value is linked to the probability of being hired or to the expected salary.

If the firm is not able to verify either dimension, then clearly no useful information about the total value $x + y$ can be credibly transmitted in equilibrium. At the other extreme, if the firm can test both skills, it can learn the candidate's type perfectly. Our question is what can happen if the firm can test just one skill.

Suppose first that the firm chooses in advance which skill to test. If it chooses math, then it learns the value x precisely, but does not get any information about y. Conversely, if it chooses coding, it learns y but gets no update on x.

It is easy to see that the firm can improve by asking the candidate to choose which test he would like to take. The candidate who is better at math ($x > y$) would then ask to take the math test, and the candidate better at coding would ask for the coding test. Then, after giving the math test to the candidate who chose it, the firm not only learns x, but also some information on y, namely, that y is distributed on $[0, x]$, and similarly the firm that ends up giving the coding test to the candidate learns something about his math skills. (Notice that it is incentive compatible for the candidate to report his best dimension: the firm's posterior expectation of $x+y$ would be $\frac{3}{2}x$ if he asks for the math test and $\frac{3}{2}y$ if he asks for the coding test, so indeed he prefers the former if and only if $x > y$.)

Is there any way the firm can learn even more?

The answer may be a surprise: the firm can learn everything, by using a randomized mechanism. This can be achieved as follows. The firm asks the candidate to report $p = \frac{x}{x+y}$, and then proceeds by giving the candidate the math test with probability p and the coding test with probability $1-p$. If the candidate plays along, then the firm will indeed achieve full learning: after giving the math test (which is possible only if $p > 0$) and observing x, it would infer y as $y = \frac{1-p}{p}x$; similarly, after giving the coding test and observing y, it would infer x as $x = \frac{p}{1-p}y$. It therefore remains to verify that it is incentive compatible for the candidate to report $p = \frac{x}{x+y}$ truthfully.

A candidate that reports $p = \frac{x}{x+y}$ truthfully makes the firm learn his true $x + y$. A candidate that deviates and reports $\hat{p}$ instead makes the firm believe that $\widehat{x+y} = x + \frac{1-\hat{p}}{\hat{p}}x$ if he gets the math test, and that $\widehat{x+y} = \frac{\hat{p}}{1-\hat{p}}y + y$ if he gets the coding test. Since he gets the former with probability $\hat{p}$ and the latter with probability $1 - \hat{p}$, in expectation he makes the impression

$$\hat{p}\left(x + \frac{1-\hat{p}}{\hat{p}}x\right) + (1-\hat{p})\left(\frac{\hat{p}}{1-\hat{p}}y + y\right) = x + y.$$

This means that the candidate cannot gain by misreporting, and indeed the firm can learn everything if it gives the candidate the freedom to choose the testing probabilities.

In what follows, we study how far the simple logic of this example generalizes. We build a model with two economic agents, a sender (he) and a receiver (she), which we can think of as a job candidate and an interviewer. The sender has multidimensional private information (e.g., his skills) and can send a message to the receiver, who can subsequently verify the value of one of the dimensions. We think of the receiver's problem as one of mechanism design: she commits to a verification rule so that equilibrium play in the resulting communication game will reveal as much information about the private type as possible. We assume that while the receiver is free to design the verification rule, she has no control over the subsequent (unmodeled) actions that will generate payoffs for the sender. The sender, in his turn, tries to maximize the overall impression of the receiver, e.g., her posterior belief about the sum of his skills.

The sender's gain from convincing the receiver that his type is a is modeled by an exogenous function $V(a)$ (in Example 1, this is the sum of coordinates). We can think of this function as a reduced-form way of modeling the outcome of any subsequent interaction between the sender and receiver. We study how the possibility or impossibility of perfect learning depends on the function $V(a)$, and we give a complete characterization of the functions $V(a)$ for which full learning is possible. In particular, when $V(a)$ is submodular, full learning is possible, whereas if $V(a)$ is strictly supermodular then it is impossible. (In the boundary case where V is additively separable, as in Example 1 above, the mechanism is essentially unique.) Our general argument uses direct-revelation mechanisms, but when $V(a)$ is submodular and satisfies some additional regularity conditions, we can also construct an indirect mechanism in which the sender chooses probabilities of testing each dimension, generalizing Example 1.

Our assumption that the receiver has no control over payoffs (for given beliefs) is natural for many settings: e.g., in the case of an interviewer and a job candidate, the interviewer might be obliged to write a truthful report of what she learned to her supervisor, so she may exercise control over what she chooses to learn, but she cannot manipulate the candidate's payoffs in any other way. Our other central assumption, that the receiver can verify exactly one dimension, is of course more stylized. We adopt this assumption to achieve the starkest results, showing that a minimal amount of verification allows full learning in Example 1; by maintaining the assumption throughout, we can ask how far the example can be pushed, in a way that allows for a crisp answer (Proposition 2). It also allows us to best connect to existing literature, as discussed below.

Our paper contributes to the large literature on strategic information transmission and communication that starts with Crawford and Sobel (1982) and Holmström (1977), and more specifically to transmission of multidimensional information (see Sobel, 2013, for an extensive review of the literature on strategic communication). In the cheap talk framework where no information is verifiable, Chakraborty and Harbaugh (2007) show that some information, in particular, relative statements about the dimensions of interest, may be transmitted. Chakraborty and Harbaugh (2010) further show that in the linear case, even when the sender's preferences are independent of his type, information on all but one dimension (the "dimensions of agreement") may be transmitted. This result has some resemblance to our example above, where one might view the verification as filling in the missing dimension. Lipnowski and Ravid (2017) consider a more general, abstract formulation and characterize optimal equilibrium outcomes for the sender. Battaglini

(2002) studies cheap talk with multiple senders; his model shares with ours the possibility of full learning.[1]

The paper that is the most related to ours is Glazer and Rubinstein (2004), which also studies a receiver ("listener") who is trying to elicit multidimensional information from the sender ("speaker") and is able to verify at most one dimension. In that paper, the receiver uses the information learnt to make a binary decision, e.g. whether to hire the sender or not, and the sender has a constant preference over decisions, e.g. always prefers to be hired.[2] In our terms this corresponds to assuming that V can take two values. The receiver wishes to minimize the probability of a mistake. The authors characterize the optimal mechanism as a solution to a particular linear programming problem, show that it takes a fairly simple form, and show that random mechanisms may be necessary to achieve the optimum. In contrast to their paper, we consider a broader range of payoffs for the sender, but focus on the possibility of full learning, which is not discussed in Glazer and Rubinstein (2004). In their setting, if full learning were possible, it would of course be optimal.[3]

Azar and Micali (2018) also study a problem in which an agent has access to a high-dimensional vector, and the principal wishes to know the value of some function of the vector, without having the whole vector communicated. They show a result with some resemblance to ours: their principal can incentivize approximate revelation of the true value while verifying just one component. They allow the principal to design the incentives freely, in contrast to our exogenously given $V(a)$.

Our paper is also related in spirit to other literature on communication with verification. Dziuda and Salas (2018) study a cheap-talk model in which the receiver may learn that the sender lied, but without learning what the truth was; in their model, discovery of lies is random and exogenous, unlike ours where verification is the object of design. Deb and Stewart (2018) study an adaptive testing problem where there is a limit on the number of tests that may be performed, as in our model. There is also a growing branch of the mechanism design literature with costly verification, started by Townsend (1979), and more recently including Kartik and Tercieux (2012), Ben-Porath, Dekel, and Lipman (2014), and Erlanson and Kleiner (2017).

The rest of the paper proceeds as follows. In Section 2, we set up the framework and

[1] Other papers addressing full or nearly-full learning with multiple senders include Ambrus and Takahashi (2008), Meyer, Moreno de Barreda, and Nafziger (2019), and Ambrus and Lu (2014).

[2] Glazer and Rubinstein (2004) also mention a number of further examples of applications, which could apply to our paper as well.

[3] Other papers studying communication of multidimensional information include Austen-Smith and Fryer (2005), Polborn and Yi (2006), and Egorov (2015).

define the notion of a valid mechanism. Section 3 analyzes the model, characterizing when full learning is possible. Section 4 considers robustness to several variations, including adding a condition on off-path beliefs that limits the possibility of punishing deviations by the sender. Section 5 is a conclusion, where we discuss some practical takeaways and open questions for further inquiry.

2 Setup

There are two agents, whom we call the *sender* and the *receiver*. The sender has multi-dimensional private information, which we call his *type* and denote $a = (a_1, \ldots, a_n) \in A$, where $A = [0, \infty)^n$ is the space of possible types. This type follows a prior distribution $\Phi \in \Delta(A)$.

After the sender and receiver interact, the receiver will be left with some (possibly probabilistic) posterior belief $\mu \in \Delta(A)$ concerning the sender's type. We take as given a function $V : \Delta(A) \to \mathbb{R}$; $V(\mu)$ denotes the payoff that the sender gets if he induces belief μ. In particular, for a type $a \in A$, we write $V(a)$ for the payoff that the sender gets if he induces a belief that is a point mass on a. For instance, in the job candidate example, $V(\mu)$ could represent the salary that the candidate will receive if the interviewer's posterior belief is μ (perhaps this is simply the posterior expectation of his marginal product for the firm). In the prosecution example, $V(\mu)$ would denote the probability that the prosecutor drops the case. More generally, we have in mind a signaling-game-like situation in which, after learning, the receiver takes some action that generates a payoff for the sender. We have no need to model this action explicitly, so instead we summarize it with the function $V(\cdot)$.

When the sender communicates with the receiver, he faces uncertainty over what belief μ will be induced: in particular, if the receiver plans to verify a randomly chosen dimension, μ may depend on which dimension is verified. We assume that $V(\cdot)$ is a von Neumann-Morgenstern utility function, so that the sender acts to maximize the expectation of $V(\mu)$. We assume throughout that V is weakly increasing: if $\mu, \mu' \in \Delta(A)$, and μ first-order stochastically dominates[4] μ', then $V(\mu) \geq V(\mu')$. We also normalize $V(0) = 0$ (hereinafter, we use 0 to denote the null vector when it does not cause confusion).

The sender and the receiver can engage in a strategic interaction with the following structure: The sender can transmit a message. The receiver can then verify one component

[4] Probability distribution μ (weakly) first-order stochastically dominates μ' if $\mathbb{E}_{x\sim\mu}[f(x)] \geq \mathbb{E}_{x\sim\mu'}[f(x)]$ for all increasing functions f for which the expectations are defined (see, e.g., Van Zandt and Vives, 2007, p. 349).

of the sender's type. We assume that if the receiver chooses to verify dimension i, she then learns the value of a_i perfectly. The receiver can commit in advance to the verification strategy, but has no control over the post-verification interaction and thus simply takes as given the function $V(\cdot)$. We also assume, in line with the mechanism design tradition, that the receiver can choose an equilibrium of the ensuing game.

We assume that the receiver simply wishes to learn as much as possible. In particular, our interest is whether there exists a way for the receiver to learn the sender's type perfectly in equilibrium. Notice that we can ask this question without committing to a particular numerical objective for the receiver to maximize.

Formally, the object chosen by the receiver — describing both the game in which the sender and receiver interact, and the equilibrium thereof — is a *mechanism*, a tuple $\mathcal{M} = (M, \sigma, p, \mu)$, where:

- M is a message space;
- $\sigma : A \to \Delta(M)$ is a (possibly mixed) reporting strategy for the sender;
- p is a (possibly mixed) verification strategy for the receiver, specifying probabilities $(p_1(m), \ldots, p_n(m))$ that sum to 1, for each $m \in M$ (here $p_i(m)$ is the probability of verifying dimension i);[5]
- μ is a belief system for the receiver, specifying posteriors $\mu(h) \in \Delta(A)$ for each $h \in H$, where

  $$H = \{(m, i, s) : m \in M, p_i(m) > 0, s \in [0, \infty)\}$$

  is the set of *(receiver) histories* that are possible given verification strategy p.

Note that the receiver's beliefs are defined as functions of the history; a history (m, i, s) means that message m was sent, dimension i was verified, and the value observed was s. We assume that once the belief μ is induced, the sender receives a payoff equal to $V(\mu)$.

We say that the mechanism is a *direct mechanism* if the sender just reports his type truthfully: $M = A$, and $\sigma(a) = a$ (deterministically) for each a.

We say that the mechanism is *valid* if the sender's strategy and beliefs constitute an equilibrium (more specifically, a weak PBE). That is, validity requires the following:

[5] Notice that this formulation *requires* the receiver to check one dimension. We could also allow for some probability $p_0(m)$ of not checking anything, at the cost of some notational inconvenience. This would not help the receiver: for any full-learning mechanism that places positive probability on no verification, we could move this probability mass onto verifying dimension 1 without weakening the incentives for truthful reporting.

- *Incentive compatibility (for sender)*: For each $a \in A$, $\sigma(a)$ has its support contained in the set of $m \in M$ that maximize

$$\sum_{i=1}^{n} p_i(m) V(\mu(m,i,a_i)).$$

- *Bayesian updating*: Let $\bar{H} = A \times H$ denote the set of *full histories*, specifying both the sender's true type and the interaction with the receiver. The prior Φ and the strategies σ, p together induce a probability distribution $\bar{\zeta}$ over $\bar{H}$. Let ζ be the marginal distribution over H. Then, we require that for any measurable set of receiver histories $H' \subset H$ and any measurable set of types $A' \subset A$,

$$\int_{H'} \mu(h)(A')\, d\zeta(h) = \bar{\zeta}(A' \times H').$$

Note that we have not required incentive compatibility for the receiver's verification. This reflects the assumption that the receiver commits in advance to the verification strategy. Our definition of H in the specification of beliefs also reflects this assumption: we do not require beliefs to be defined at "histories" (m, i, s) that would be reachable only if the receiver failed to follow the verification strategy.

Bayesian updating serves to pin down beliefs at on-path histories, but imposes no constraints on beliefs at off-path histories. It will sometimes be convenient to focus on mechanisms satisfying the following:

- *Punishment beliefs*: For any receiver history $h = (m, i, s)$ that is outside the support of ζ, the belief $\mu(h)$ is a point mass on type 0.

Indeed, by a standard argument, any outcome that can be supported by some valid mechanism can in particular be supported by one with punishment beliefs.[6]

Notice, however, that punishment beliefs effectively mean that at off-path histories, the receiver does not place full faith in the accuracy of the verification technology, since if $s \neq 0$, the verification shows that the sender is not actually the zero type. We could alternatively impose the following condition:

- *Trusted verification*: For any history $h = (m, i, s)$, the belief $\mu(h)$ puts no probability on types a with $a_i \neq s$.

[6] See the proof of Lemma 0 in the Appendix.

For our main analysis, we will not impose this condition. The model without trusted verification has several interpretations: We could view the model as a limiting case where the receiver has infinitesimal uncertainty about the correctness of the verification technology. We could also treat it as a shorthand for a situation in which the receiver can commit to give the worst payoff $V(0)$ when the sender is known to have deviated (for example, in the employment application, we might simply imagine that the company refuses to hire a candidate who has been caught lying; in the insurance claim example, the insurance company might not be obligated to honor any claims if it has shown that one claim was false). Finally, we can also associate it with an alternative model in which the receiver can only perform verifications of the form "is a_i equal to s?" for a specific value of s, rather than "what is a_i?" In such a model, an off-path negative answer would generally not preclude the punishment belief that places all weight on type 0. (For brevity, we avoid writing out this alternative model in full.) In any case, in Section 4 we will consider imposing trusted verification and will show that our main conclusions are robust to it.

We are particularly interested in mechanisms that allow full learning of the sender's type.[7]

- *Full learning*: For every type a, at every history $h \in H(a \mid \mathcal{M})$, the belief $\mu(h)$ is a point mass on type a. Here, we define

  $$H(a \mid \mathcal{M}) = \{(m, i, a_i) \in H : m \in \operatorname{supp}(\sigma(a))\},$$

  the set of histories that can arise when the sender has type a.

Before moving on, we add one observation: The function $V(\cdot)$ appears only in the incentive compatibility condition, and this condition is invariant under translating the whole function $V(\cdot)$ by a constant. Thus it is indeed just a normalization to assume that $V(0) = 0$.

[7] In some applications, we may think the receiver is content to learn the value of $V(a)$ without learning a itself: e.g. the employer may be interested in knowing the worker's total output, but not how it is achieved. While learning $V(a)$ may appear to be a simpler problem than learning a, in fact it is not: if there is a valid mechanism that allows the receiver to learn $V(a)$, there is also one that achieves full learning. For a formal statement and proof, see Proposition A2 in the Appendix.

3 Main Analysis

3.1 Initial observations

Our main question is: For what payoff functions $V(\cdot)$ does there exist a mechanism that achieves full learning?

We begin with a version of the revelation principle. This shows that we can restrict attention to direct mechanisms, and also can assume punishment beliefs as described above.

Lemma 0. *If there exists a valid mechanism with full learning, then there exists a valid direct mechanism satisfying punishment beliefs and full learning.*

The proofs of this and other results that are not given in the text are in the Appendix.

By focusing on direct mechanisms that furthermore satisfy punishment beliefs, we see that we need only specify the verification strategy p, since the message space, sender strategy and beliefs are pinned down. Specifically, full learning is possible if and only if there exists a choice of verification probabilities $p = (p_1, \ldots, p_n)$, with each $p_i : A \to [0, 1]$ and $\sum_i p_i(a) = 1$ for all a, satisfying the incentive compatibility condition for all types a and $\hat{a}$:

$$V(a) \geq \left(\sum_{i:\ \hat{a}_i = a_i} p_i(\hat{a}) \right) V(\hat{a}). \tag{1}$$

Indeed, here the left side represents the payoff that the sender gets from truthfully reporting type a, which will be $V(a)$ no matter which dimension is verified, and the right side is the expected payoff from reporting $\hat{a}$, given the punishment beliefs.

As the condition (1) makes clear, we do not actually need to concern ourselves with $V(\mu)$ for arbitrary beliefs μ; only the values of V on degenerate beliefs matter. So for the remainder of the paper, we will think of V as being defined only on A, instead of on $\Delta(A)$. Then, the monotonicity requirement just says that if $a' \leq a$ (componentwise) then $V(a') \leq V(a)$.

3.2 Additively separable case

We begin by reconsidering (and slightly generalizing) Example 1 from the Introduction. We show that the verification probabilities that allow full learning not only exist for any n, but are essentially unique.

Specifically, suppose $V(a)$ is additively separable in its components, so

$$V(a)=\sum_{i=1}^{n} v_i\left(a_i\right), \tag{2}$$

where $v_i : [0,\infty) \to \mathbb{R}$ are increasing functions. Since we assumed $V(0)=0$, we may pick $v_i(\cdot)$ such that $v_i(0)=0$ for each i.

Proposition 1. *Suppose that V is additively separable and defined by (2). Then, full learning is achieved by the valid direct mechanism using the verification probabilities*

$$p_i(a)=\frac{v_i\left(a_i\right)}{V\left(a\right)}=\frac{v_i\left(a_i\right)}{v_1\left(a_1\right)+\cdots+v_n\left(a_n\right)} \quad (1\leq i\leq n)$$

for each a such that $V(a)\neq 0$ (and arbitrary verification probabilities for a such that $V(a)=0$). Furthermore, these probabilities are unique: If $\mathcal{M}=(M,\sigma,p,\mu)$ is a valid (possibly indirect) mechanism with full learning, then for any type $a\in A$ with $V(a)>0$, for any $m\in \operatorname{supp}(\sigma(a))$, we have

$$p_i(m)=\frac{v_i\left(a_i\right)}{v_1\left(a_1\right)+\cdots+v_n\left(a_n\right)} \quad (1\leq i\leq n).$$

Proof. For existence, we just need to check that incentive compatibility (1) is satisfied. For any $\hat{a}$, the right-hand side of (1) equals $\sum_{i:\hat{a}_i=a_i} v_i(\hat{a}_i)$. This is clearly at most $\sum_{i=1}^{n} v_i(a_i)=V(a)$, as needed.

To prove uniqueness, consider any type a, and the alternative type a' that agrees with a in all coordinates except in coordinate i, where $a'_i=0$. Let m be any message in the support of $\sigma(a)$.

The assumption of full learning implies that, if type a' sends message m and coordinate i is not verified, then the resulting belief places probability 1 on type a, and the sender gets reward $V(a)$. Hence, the expected payoff to sending message m is at least $(1-p_i(m))V(a)$. So incentive compatibility for the pair of types a' and a implies

$$V(a')\geq (1-p_i(m))V(a),$$

which implies

$$p_i(m)\geq \frac{V(a)-V(a')}{V(a)}=\frac{v_i(a_i)}{v_1(a_1)+\cdots+v_n(a_n)}.$$

Since we must also have $\sum_{i=1}^{n} p_i(m)=1$, these inequalities must hold as equalities. □

The mechanism suggested in Proposition 1 has several remarkable properties. To state them, assume for simplicity that each v_i is strictly increasing and continuous, and in particular $V(a) = 0 \Leftrightarrow a = (0, \ldots, 0)$.

- The mechanism can be implemented as an indirect mechanism, as in the Introduction, where the sender chooses a probability distribution $(q_1, \ldots, q_n)$ over dimensions to verify (so M is an $(n-1)$-dimensional simplex of probabilities). When dimension i is verified and the observed value is s, the receiver infers $v_j(a_j) = \frac{q_j}{q_i} v_i(s)$ for each j, and so infers a completely by inverting each v_j.

- The mechanism also does not actually require the receiver to commit to the verification strategy, as we have assumed. Indeed, if she could freely choose which component to verify, note that once she has heard message m, she expects to end up believing that the sender is type m (and to give reward $V(m)$) regardless of which component she verifies, so she is indifferent at this stage.

- The mechanism does not depend on the distribution of sender's type Φ. Moreover, it would perform just as well if the receiver had a wrong belief about Φ. Implementing this mechanism therefore requires the receiver to know the payoff function $V(\cdot)$ and nothing else.

- In the case where all v_i are linear, the indirect implementation highlights that the parties do not need to agree on the "scale" in which the type is measured, i.e. it works even if the sender perceives his type as $(\lambda a_1, \ldots, \lambda a_n)$ rather than $(a_1, \ldots, a_n)$, for an arbitrary positive scalar λ.

As it turns out, all these properties, with the exception of the last one, hold quite a bit more generally.

3.3 General characterization

We now provide a necessary and sufficient condition for full learning to be achievable. For this we need a bit of notation. Whenever $S \subset \{1, \ldots, n\}$ is a set of indices and $a \in A$, define $a|_S$ as the type that agrees with a on the components $i \in S$, and whose other coordinates are all zero. Also, when S has a single element i, we will write $a|_i$ rather than $a|_{\{i\}}$.

Proposition 2. *There exists a valid mechanism that achieves full learning if and only if V satisfies the following condition. For every $a \in A$, and any collection of nonnegative* *weights λ_S for each of the 2^n sets $S \subset \{1, \dots, n\}$ that satisfies $\sum_{S:i\in S} \lambda_S = 1$ for each index $i = 1, \dots, n$, we have*

$$V(a) \leq \sum_{S\subset\{1,\dots,n\}} \lambda_S V(a|_S).$$

To see why the characterization takes this form, consider what happens to the incentive condition (1) when we hold fixed the report $\hat{a}$, and also hold fixed the coordinates a_i of the true type for which $a_i = \hat{a}_i$, but vary the other coordinates a_j. Then the right side of (1) is constant, while the left side is increasing in a. Consequently, the constraint is tightest when $a = \hat{a}|_S$ for some set S: if we can deter these types a from reporting $\hat{a}$, then all other types are deterred as well. So full learning is achievable as long as we can choose the verification probabilities for each type a to deter misreporting by the (finitely many) types $a|_S$. The proposition gives a duality-based characterization of when this is possible.

The condition in Proposition 2 takes a particularly simple form if $n = 2$: the only possible weights are of the form $\lambda_{\{1\}} = \lambda_{\{2\}} = \lambda$ and $\lambda_{\{1,2\}} = 1 - \lambda$, and the condition simplifies to $V(a) \leq V(a|_1) + V(a|_2)$. Indeed, in this case, the argument from the previous paragraph implies that taking $p_1(a) = V(a|_1)/V(a)$ and $p_2(a) = 1 - p_1(a)$ will suffice. The following result describes the complete set of verification probabilities.

Proposition 3. *If $n = 2$, then there is a valid mechanism that achieves full learning if and only if $V(a) \leq V(a|_1) + V(a|_2)$ for each a. A direct mechanism is valid if and only if verification probabilities satisfy, for any a:*

$$p_1(a) \geq 1 - \frac{V(a|_2)}{V(a)}, \qquad p_2(a) \geq 1 - \frac{V(a|_1)}{V(a)}.$$

3.4 Submodular and supermodular functions

Here we give a couple of illustrative applications of Proposition 2.

First, suppose that the payoff function V is submodular.[8] In this case, the condition in Proposition 2 holds, and in fact the proof of that proposition leads to a simple explicit construction for a direct mechanism with full learning, which we state as a separate result.

To state the result formally, extending our $a|_S$ notation, for each $a \in A$ and each $i = 1, \dots, n$, let $a|_{[i]}$ be the type whose first i components agree with a, and whose remaining $n - i$ components are all zero. Consistently with this, let also $a|_{[0]} = (0, \dots, 0)$.

[8] The function V is *submodular* if $V(a \vee a') + V(a \wedge a') \leq V(a) + V(a')$ for all a, a', where $\vee$ denotes componentwise max and $\wedge$ denotes componentwise min. V is *strictly submodular* if the inequality holds strictly whenever $\{a \vee a', a \wedge a'\} \neq \{a, a'\}$. V is *supermodular* (resp. strictly supermodular) if $-V$ is submodular (resp. strictly submodular). Additively separable functions are both sub- and supermodular.

Proposition 4. *Suppose that V is submodular. Then the following valid direct mechanism achieves full learning: If $V(a) > 0$, dimension i is verified with probability*

$$p_i(a) = \frac{V(a|_{[i]}) - V(a|_{[i-1]})}{V(a)},$$

and if $V(a) = 0$, the probabilities are chosen arbitrarily.

On the other hand, if V is (strictly) supermodular, full learning is not achievable. For example, suppose $V(a_1, a_2) = \min\{a_1, a_2\}$. To deter deviations to reporting type $(1, 1)$, the first dimension must be tested with probability 1 (otherwise type $(0, 1)$ would misreport), but likewise the second dimension must be tested with probability 1, and we cannot do both. By the exact same reasoning, $V(a_1, a_2) = a_1a_2$ would not allow full learning either.

In fact, we can give a broader impossibility result:

Proposition 5. *Suppose that there is a type a such that*

$$V(a) > \sum_{i=1}^{n} V(a|_i). \tag{3}$$

Then there does not exist a valid mechanism that achieves full learning.

Note that if V is strictly supermodular (and $V(0) = 0$ as we have assumed), then the condition in the proposition is satisfied for *any* type a that is positive in every coordinate. So, the proposition covers such functions (but is also much more general).

Proof. Take type a for which the inequality holds. Note that the condition in Proposition 2 is violated, by taking $\lambda_{\{i\}} = 1$ for each i, and $\lambda_S = 0$ for all non-singleton sets. □

For a simple intuition about why the submodular versus supermodular distinction arises, think about the job candidate with two possible skills, as in Example 1. A candidate who is strong on one skill but weak on the other has a potential incentive to pretend to be strong on both. This can be deterred if the weak skill is verified with sufficiently high probability. But if the skills are complements (supermodular case), the gains from appearing to be strong on both skills rather than just one are high, and there is no way to choose verification probabilities to deter both a (strong math, weak coding) candidate and a (weak math, strong coding) candidate. Whereas if the skills are substitutes (submodular case), the gains are smaller and this can be done.

For some specific, stark examples, consider first the function $V(a_1, a_2) = \max\{a_1, a_2\}$, which is submodular. In this case, full learning is achievable simply by always testing whichever coordinate is reported higher — which is in line with Proposition 3. In contrast, in the case of function $V(a_1, a_2) = \min\{a_1, a_2\}$, which is supermodular, full learning is not achievable.

3.5 Indirect mechanisms

The results presented in the preceding subsections provide a general characterization of when full learning is achievable. The construction employed a direct mechanism that, in particular, required punishing the sender with the worst possible belief in case verification failed.

In Example 1, however, we used an indirect mechanism, in which the sender effectively just reports the probability vector q by which he should be tested, and the one verified coordinate is then used to infer all other coordinates. In general, this has a few advantages. First, essentially all histories are on-path (provided that Φ has full support), so we do not need to worry about the choice of off-path beliefs. Second, direct mechanisms with punishment beliefs are fragile in the sense that, if the sender's belief about his own type is off by an ε amount, the receiver ends up with a posterior that is very far from the truth; indirect mechanisms avoid this fragility, as long as V is continuous.[9] A third advantage is that, if the receiver could actually choose not to verify anything, and verifying came at a small cost $\varepsilon > 0$, then in a direct mechanism, the receiver ex-post would not have the incentive to actually carry out the verification, whereas in an indirect mechanism she could, since she is still uncertain about the type after hearing the message.

Hence, we might naturally wonder whether the indirect mechanism can be readily generalized to other $V(\cdot)$. As it turns out, it generalizes quite broadly: we can give a construction for any V that is submodular and satisfies some regularity conditions, although our construction is not quite as explicit as the one in Proposition 4 above.

Specifically, suppose that V is submodular and continuously differentiable, and write V_i for the derivative with respect to coordinate i. Suppose further that all partial derivatives V_i are bounded in an interval $[k, K]$, where $0 < k < K < \infty$. These assumptions will be maintained for the remainder of this subsection. Note that the set of submodular functions satisfying these regularity conditions is dense in the set of all increasing continuous

[9] In the working paper version, Carroll and Egorov (2018), we also consider a related fragility issue: what if the signal is noisy, instead of revealing a_i exactly? Then direct mechanisms fail, and the receiver must resort to indirect mechanisms. We construct a parameterized example to show that the receiver can use indirect mechanisms to learn the sender's type almost perfectly when the level of noise is small.

submodular functions, in the topology of uniform convergence on compact sets.

For any vector $q = (q_1, \dots, q_n)$ of probabilities summing to 1, define a parametric curve $a(q,t) = (a_1(q,t), \dots, a_n(q,t))$ for $t \geq 0$ by the differential equations

$$\frac{\partial a_i}{\partial t} = \frac{q_i}{V_i(a(q,t))} \tag{4}$$

and the initial condition $a(q,0) = 0$. In the indirect mechanism, we simply have the agent report the probability vector q for the curve his type lies on, and the receiver verifies each dimension i with the corresponding probability q_i.

Of course, for this mechanism to be well-defined, we need to know that every possible type does indeed lie on some such curve.

Lemma 6. *For every type $a \in A$, there exist q and t such that $a(q,t) = a$.*

Note that in fact, a lies on the curve defined by q if and only if $a = a(q, V(a))$. This follows from the fact that $\frac{\partial}{\partial t} V(a(q,t)) = \sum_i \frac{\partial a_i}{\partial t} \cdot V_i(a(q,t)) = \sum_i q_i = 1$, hence $V(a(q,t)) = t$ for all t.

We have not ruled out the possibility that the type a lies on more than one such curve.[10] (And of course this is true for $a = 0$, which lies on every curve.) In this case, we will have type a mix according to an arbitrary full-support distribution over the relevant set of curves.

If i is a coordinate such that $q_i > 0$, notice that (4), together with our bounds on derivatives, ensures that $a_i(q,t)$ is strictly increasing and goes to ∞ as $t \to \infty$. Continuity then implies that for every $s \geq 0$, there exists a unique t such that $a_i(q,t) = s$. Type $a(q,t)$ can generate the history (q,i,s) by reporting as prescribed above; thus every history in H is on-path.

In summary, the mechanism is described as follows:

- The message space consists of all probability vectors $q = (q_1, \dots, q_n)$, with $q_i \geq 0$ and $\sum_i q_i = 1$.
- For the reporting strategy, each type a uses an (arbitrary) full-support distribution over the set of q such that a lies on the curve defined by q. (This set is nonempty, by Lemma 6, and closed since it is given by the equation $a = a(q, V(a))$.)
- Given message q, the receiver verifies each coordinate i with probability q_i.

[10] For $n = 2$, it is easy to show that the curves do not intersect except at $a = 0$.

- At any history $(q,i,s) \in H$, the receiver's belief puts probability 1 on $a(q,t)$ where t is the unique value satisfying $a_i(q,t) = s$.

Proposition 7. *Suppose that V is submodular and continuously differentiable, and all partial derivatives V_i are bounded in the interval $[k,K]$. Then the indirect mechanism described above is a valid mechanism that achieves full learning.*

The proof involves considering a deviation to a report q by a type not on the $a(q,\cdot)$ curve, and making judicious use of submodularity to compare the payoff from deviation against the equilibrium payoff and show that the deviation cannot be beneficial.

4 Extensions

4.1 Trusted verification

As mentioned in Section 2, it is natural to consider imposing the trusted verification condition as a restriction on beliefs when the sender is found to have misreported. How much do our results change under this restriction?

First, our major qualitative conclusions remain unchanged. In particular, full learning is still possible whenever $V(\cdot)$ is submodular, although the explicit mechanism from Proposition 4 no longer works,[11] and indeed, we do not know of a similarly simple explicit formula for a mechanism that works in general. Actually, when V satisfies the regularity assumptions of Proposition 7, that proposition already shows that full learning is possible; notice that trusted verification is automatically satisfied since every history $h = (m,i,s)$ is on-path. But even for submodular functions that fail those regularity assumptions, full learning is possible:

Proposition 8. *Suppose that V is submodular. Then there exists a valid direct mechanism that achieves full learning and satisfies trusted verification.*

The proof of Proposition 8 is nonconstructive. As before, the idea is to find verification probabilities $p(a)$ for any fixed $a \in A$ that deter any other type $z \neq a$ from deviating by reporting type a, and we now use the Kakutani fixed-point theorem to show that such probabilities exist. More specifically, for any verification probability vector p, we consider the set of types $z \leq a$ that would gain the most from misreporting as a. We then let E_p be

[11] For example, take $n = 2$ and $V(a_1,a_2) = \max\{a_1,a_2\}$. Then under the verification strategy from Proposition 4, type $(1,1)$ can strictly gain from reporting as type $(1,2)$ if the receiver's beliefs upon detecting the lie are constrained by trusted verification.

the set of all alternative verification probability vectors that would successfully deter these types from deviating. This set is quickly shown to be nonempty using the submodularity of V. It turns out that the correspondence $p \mapsto E_p$ is not upper-hemicontinuous, so we cannot apply the Kakutani fixed-point theorem immediately, but we can "smooth out" E_p appropriately to yield a correspondence for which the theorem does apply. Taking p to be a fixed point, then, all types that would gain the most from deviating under p are deterred from deviating, which is exactly what we need.

Our other main conclusion from Section 3 was that strictly supermodular V does not allow full learning (Proposition 5). Clearly this conclusion also still holds up when we restrict mechanisms by requiring trusted verification.

With trusted verification, we do not know of a complete characterization of the functions $V(\cdot)$ for which full learning is possible. However, we do have such a characterization for the two-dimensional case:

Proposition 9. *Suppose that $n = 2$. Then full learning is achievable with a valid mechanism satisfying trusted verification if and only if V satisfies the following property: for any two types $x, a \in A$ with $x < a$, we have*

$$(V(a) - V(x_1, a_2))(V(a) - V(a_1, x_2)) \leq (V(x_1, a_2) - V(x|_1))(V(a_1, x_2) - V(x|_2)).$$

The proof in fact gives an explicit construction of a full-learning mechanism when the condition is satisfied. One can also use this result to construct functions for which full learning is possible without the trusted verification requirement, but not possible with it, thus showing that the restriction on beliefs does have bite.

4.2 Concave transformations

Another interesting property is that any mechanism that achieves full learning is robust to concave transformations of the sender's payoff function:

Proposition 10. *Let V be such that full learning is achievable in a valid direct mechanism $\mathcal{M}$. Then the same mechanism $\mathcal{M}$ also remains valid when the payoff function is $V' = U \circ V$, where $U : [0, \infty) \to [0, \infty)$ is any increasing, concave transformation.*

Essentially, the result holds because when a mechanism achieves full learning, the sender is certain of his payoff along the equilibrium path, whereas by deviating he gets a lottery over payoffs. Concave transformations make such a lottery even less desirable.

Concave transformations can arise naturally in two ways. First, if V is the monetary payoff that the sender receives (for example, if he is a job candidate who is paid his perceived marginal product), then U can represent risk aversion. Thus, the proposition says that any mechanism that achieves full learning for a risk-neutral sender also works when the sender is risk-averse. Second, V might represent value measured in some abstract units, and U can represent decreasing returns. For example, if the job candidate's "total skill" is $a_1 + \cdots + a_n$, the proposition says that any mechanism that works when the candidate's marginal product equals his total skill also works when there are decreasing returns to total skill.

4.3 Payoffs depending on sender's type

We have so far assumed that the sender's payoff $V(a)$ (or, more generally, $V(\mu)$) depends only on the receiver's posterior belief, but not on the sender's true type. In some natural cases, however, this assumption ought to be relaxed. To allow for these possibilities, let us write $V(a;x)$ to denote the payoff of a sender of type x if the receiver's posterior is that he is of type a. We have the following result.

Proposition 11. *Suppose that $V(a;a)$ satisfies the condition in Proposition 2 and $V(a;x) \leq \max\{V(a;a), V(x;x)\}$ for all a and x, and furthermore, $V(0;x) = 0$ for every x. Then there is a valid direct mechanism that achieves full learning.*

Proposition 11 says that the previous results generalize as long as type x pretending to be type a cannot be better off than both types x and a telling the truth. This would be the case, for example, if there is a cost of lying: in this case, $V(a;x) \leq V(a;a)$ (and in particular, $V(0,x) \leq V(0;0) = 0$) and the condition holds. But the assumption is more general than that. Suppose, for example, that to sustain his reputation, a lower type of candidate must exert extra effort after being hired, while a higher type can afford to slack. For example, let $V(a;x) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \kappa (a_i - x_i)$ for $a \neq 0$, so the cost of effort is proportionate to the difference between the reputed skill level and the true skill level, with coefficient $\kappa \in (0,1)$ (and suppose that the worst type 0 is never hired, so $V(0;x) = 0$ for all x). Then

$$\begin{aligned} V(a;x) &= \sum\nolimits_{i=1}^{n} (1-\kappa) a_i + \sum\nolimits_{i=1}^{n} \kappa x_i \\ &= (1-\kappa) V(a;a) + \kappa V(x;x) \\ &\leq \max\{V(a;a), V(x;x)\}, \end{aligned}$$

so the condition is satisfied. This Proposition also shows the limits of the argument: for example, if $\kappa > 1$, then for a high type the temptation to pretend to be a low type and save on effort would be too high; in that case, clearly, full learning would not be feasible.

5 Conclusion

We considered the problem of strategic transmission of multidimensional information between a sender and a receiver, where the receiver is able to verify at most one dimension. If the receiver chooses this dimension without any input from the sender, she learns just that dimension, at least if dimensions are uncorrelated. An obvious improvement is to ask the sender which dimension to test; in this case, the receiver perfectly learns that dimension, and the sender's choice reveals some information about the other dimensions as well. The main contribution of our paper is showing that if we take this logic just one step further and allow for randomizations over tests, the receiver may learn the sender's type fully, for a wide range of the sender's objective functions. While the main focus of the theoretical results has been on direct mechanisms, we also showed that in the main leading cases, full learning is possible using an indirect mechanism in which the sender just chooses the probability of testing each dimension.

The paper's main contribution is theoretical, but we believe it has practical takeaways. In our view, the indirect mechanism, where the sender suggests probabilities to verify each dimension, is not so far from the structure of interactions that may occur in practice. For example, it is quite common for an interviewer to ask a job candidate to describe a project (or, in an academic context, a paper) that he listed on the vita, with the understanding that the candidate will proceed with the best one. But the candidate may instead offer the interviewer to make the selection, or suggest a couple to choose from, or he may even suggest a few and try to nudge the interviewer towards one or the other. Clearly, this communicates additional information about the candidate's willingness to talk about each project, which is very much in line with the spirit of the proposed mechanism. Similarly, the idea of drawing inference from choice of tests has recently gained attention in the insurance literature, see Crocker and Zhu (2018).

The paper's results suggest many interesting directions for further inquiry. One question is how much the receiver can learn when full learning is impossible. In the working paper version, Carroll and Egorov (2018), we show how a construction building on the ideas in this paper can be used to fully separate an arbitrarily large set of types when V is "close" to submodular, but this does not imply that doing so is optimal. For another direction, suppose that even offering one test is costly (as in e.g. Ben-Porath, Dekel, and Lipman, 2014). In this case, if full learning is achievable, it might not be optimal, since the receiver could economize by not testing over a small range of types close to zero; again, it is natural to wonder about the structure of an optimal mechanism. As another possible application, consider a professor who wants to test her students on multiple topics. In this example, running our proposed mechanism would consist of asking students to report their relative skills and then administering a test with just one (randomly determined) question. This might not be desirable, either because any single question reveals too noisy a signal, or because the students may not know their relative skills perfectly. Here, the natural solution is to offer several problems instead of one, which in turn poses the problem of the optimal number of questions an exam should have, and how to choose their topics for each student in an optimal way (see also Deb and Stewart, 2018, who study a similar question with a one-dimensional type space). We hope that the insights from this paper will help address these and other questions.

References

[1] Ambrus, Attila, and Shih En Lu (2014), “Almost fully revealing cheap talk with imperfectly informed senders,” *Games and Economic Behavior*, 88: 174–189.

[2] Ambrus, Attila, and Satoru Takahashi (2008), “Multi-sender cheap talk with restricted state spaces,” *Theoretical Economics*, 3(1): 1-27.

[3] Austen-Smith, David, and Roland G. Fryer Jr. (2005), “An Economic Analysis of ‘Acting White,’” *Quarterly Journal of Economics*, 120(2): 551–583.

[4] Azar, Pablo, and Silvio Micali (2018), “Computational Principal-Agent Problems,” *Theoretical Economics*, 13(2): 553–578.

[5] Battaglini, Marco (2002), “Multiple Referrals and Multidimensional Cheap Talk,” *Econometrica*, 70(4): 1379–1401.

[6] Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman (2014), “Optimal allocation with costly verification,” *American Economic Review*, 104(12): 3779–3813.

[7] Carroll, Gabriel, and Georgy Egorov (2018), “Strategic Communication with Minimal Verification,” unpublished paper.

[8] Chakraborty, Archishman, and Rick Harbaugh (2007), “Comparative Cheap Talk,” *Journal of Economic Theory*, 132(1): 70–94.

[9] Chakraborty, Archishman, and Rick Harbaugh (2010) “Persuasion by Cheap Talk,” *American Economic Review*, 100(5): 2361–2382.

[10] Crocker, Keith, and Nan Zhu (2018), “The Efficiency of Voluntary Risk Classification in Insurance Markets,” unpublished paper.

[11] Deb, Rahul, and Colin Stewart (2018), “Optimal Adaptive Testing: Informativeness and Incentives,” *Theoretical Economics*, 13(3): 1233–1274.

[12] Dziuda, Wioletta, and Christian Salas (2018), “Communication with Detectable Deceit,” unpublished paper.

[13] Egorov, Georgy (2015), “Single-issue Campaigns and Multidimensional Politics,” NBER Working Paper No. w21265.

[14] Erlanson, Albin, and Andreas Kleiner (2017), “Costly Verification in Collective Decisions,” unpublished paper.

[15] Glazer, Jacob, and Ariel Rubinstein (2004), "On Optimal Rules of Persuasion," *Econometrica*, 72(6): 1715–1736.

[16] Holmström, Bengt (1977), "On Incentives and Control in Organizations," Ph.D. Thesis, Stanford University.

[17] Jamison, Robert E., and William H. Ruckle (1976), "Factoring Absolutely Convergent Series," *Mathematische Annalen*, 224(2): 143–148.

[18] Kartik, Navin, and Olivier Tercieux (2012), "Implementation with Evidence," *Theoretical Economics*, 7(2): 323–355.

[19] Lipnowski, Elliot, and Doron Ravid (2017), "Cheap Talk with Transparent Motives," unpublished paper.

[20] Meyer, Margaret, Inés Moreno de Barreda, and Julia Nafziger (2019), "Robustness of Full Revelation in Multisender Cheap Talk," *Theoretical Economics*, forthcoming.

[21] Polborn, Mattias K., and David T. Yi (2006), "Informative Positive and Negative Campaigning," *Quarterly Journal of Political Science*, 1(4): 351–371.

[22] Sobel, Joel (2013), "Giving and Receiving Advice," In Advances in Economics and Econometrics, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel. Vol. 1. (Cambridge: Cambridge University Press): 305–341.

[23] Townsend, Robert (1979), "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, 21(2): 265–293.

[24] Van Zandt, Timothy, and Xavier Vives (2007), "Monotone equilibria in Bayesian games of strategic complementarities," *Journal of Economic Theory*, 134(1): 339–360.

A Appendix

Proof of Lemma 0. Let $\mathcal{M} = (M, \sigma, p, \mu)$ be a valid mechanism that achieves full learning. We wish to construct p', μ' that (together with the message space $M' = A$ and the truthful reporting strategy $\sigma'(a) = a$) form a valid direct mechanism $\mathcal{M}'$ that achieves full learning. Let $p'_i(a) = \mathbb{E}_{m \sim \sigma(a)}[p_i(m)]$, the expected probability with which dimension i is verified for type a in the original mechanism. Beliefs μ' are uniquely determined by the criteria of full learning and punishment beliefs: at histories (a, i, a_i) that can be generated by truthful reporting, the belief is degenerate on type a; at other histories (a, i, s), it puts probability 1 on type 0.

It is immediate from the construction that the mechanism satisfies full learning and punishment beliefs. To see that the mechanism is valid, we check the two conditions. For incentive compatibility, notice that if type a reports truthfully he gets a payoff of $V(a)$, whereas by reporting $\hat{a}$ he gets a payoff

$$\begin{aligned}
\sum_{i=1}^{n} p'_i(\hat{a}) V(\mu'(\hat{a}, i, a_i)) &= \mathbb{E}_{m \sim \sigma(\hat{a})} \left[\sum_{i=1}^{n} p_i(m) V(\mu'(\hat{a}, i, a_i)) \right] \\
&\leq \mathbb{E}_{m \sim \sigma(\hat{a})} \left[\sum_{i=1}^{n} p_i(m) V(\mu(m, i, a_i)) \right] \\
&\leq V(a).
\end{aligned}$$

Here the first inequality follows from the fact that for each $m \in \operatorname{supp}(\sigma(\hat{a}))$ and each i, if $p_i(m) > 0$ then either $\hat{a}_i = a_i$ implying $V(\mu(m, i, a_i)) = V(\hat{a}) = V(\mu'(\hat{a}, i, a_i))$ by full learning in the original mechanism, or $\hat{a}_i \neq a_i$ and $V(\mu'(\hat{a}, i, a_i)) = V(0) = 0$ by construction. The second inequality comes from incentive compatibility of the original mechanism.

Finally, Bayesian updating is immediate, since in equilibrium, with ex ante probability 1, the receiver puts probability 1 on the true type, which equals the report. □

Proof of Proposition 2. First we show necessity. Take the weights λ_S as given; we can assume $\lambda_\varnothing = 0$, since the value of $\lambda_\varnothing$ has no effect either on the validity of the collection of weights or on the inequality to be proven. Type $a|_S$ can, by imitating type a, get at least $\left(\sum_{i \in S} p_i(a)\right) V(a)$. Hence, incentive compatibility implies

$$\left(\sum_{i \in S} p_i(a) \right) V(a) \leq V(a|_S).$$

Now multiply by λ_S, and then sum over all S. On the left side, for each $i = 1, \ldots, n$, $p_i(a)$ appears with total weight $\sum_{S:i\in S} \lambda_S = 1$. Hence, we get

$$\left(\sum_{i=1}^{n} p_i(a)\right) V(a) \leq \sum_{S} \lambda_S V(a|_S).$$

The left side is just $V(a)$, showing that the asserted condition holds.

Now we prove sufficiency. For each type a, we need to construct the appropriate verification probabilities $p_i(a)$ to discourage deviations to a. If $V(a) = 0$ we can choose these probabilities arbitrarily, as clearly no type would deviate to such a. Now assume $V(a) > 0$.

We claim that there exist nonnegative numbers $r_1, \ldots, r_n$ such that $r_1 + \cdots + r_n = V(a)$ and, for each subset $S \subset \{1, \ldots, n\}$, $\sum_{i\in S} r_i \leq V(a|_S)$.

Suppose not. Then, applying a theorem of the alternative, we get the existence of nonnegative numbers λ_S, for each $S \subset \{1, \ldots, n\}$, such that $\sum_{S:i\in S} \lambda_S \geq 1$ for each i and $\sum_S \lambda_S V(a|_S) < V(a)$.

This is almost a contradiction to our assumed condition on V, except that for each index i, the total weight on sets containing i is ≥ 1, rather than exactly 1 as required. However, if the inequality is strict, then we can take some of the weight on a set S containing i and transfer it to set $S\backslash\{i\}$. This decreases the total weight on sets containing i, without changing the total weight on sets containing j, for any $j \neq i$. Iterating this, we can eventually get the total weight on sets containing i to be exactly 1 for each i. Moreover, each such operation can only decrease the value of $\sum_S \lambda_S V(a|_S)$, since V is monotone and we are transferring weight from larger to smaller sets. Hence the final weights will satisfy $\sum_{i\in S} \lambda_S = 1$ for each index i, and will still satisfy $\sum_S \lambda_S V(a|_S) < V(a)$, thus contradicting the assumption.

This implies the desired numbers $r_1, \ldots, r_n$ exist. Define the verification probabilities by $p_i(a) = r_i/V(a)$. We just need to check incentive compatibility condition (1).

Suppose the sender has type a, but reports $\hat{a}$. Let S be the set of coordinates i for which $\hat{a}_i = a_i$. Then

$$\sum_{i\in S} p_i(\hat{a}) = \frac{\sum_{i\in S} r_i}{V(\hat{a})} \leq \frac{V(a|_S)}{V(\hat{a})} \leq \frac{V(a)}{V(\hat{a})},$$

which is exactly what (1) requires. □

Proof of Proposition 3. The fact that the given condition is necessary and sufficient to achieve full learning follows from the discussion immediately preceding the proposition statement. Similarly, this discussion shows that a direct mechanism is valid if and only if types $a|_1$ and $a|_2$ are both deterred from reporting as type a, for each a. The relevant incentive constraints are $V(a|_1) \geq p_1(a)V(a) = (1-p_2(a))V(a)$ and $V(a|_2) \geq p_2(a)V(a) = (1-p_1(a))V(a)$, which are equivalent to the conditions given in the proposition statement. □

Proof of Proposition 4. Let us prove that the proposed mechanism is incentive compatible. If the sender has type a and reports truthfully, he evidently gets $V(a)$. If he falsely reports $\hat{a}$, then he gets the reward $V(\hat{a})$ only if the verified dimension i is such that $\hat{a}_i = a_i$; let S be the set of such indices i. Using the notation $a|_S$ as in the text, the sender's expected payoff from misreporting is

$$\begin{aligned}
\sum_{i \in S} \frac{V(\hat{a}|_{[i]}) - V(\hat{a}|_{[i-1]})}{V(\hat{a})} V(\hat{a}) &= \sum_{i \in S} (V(\hat{a}|_{[i]}) - V(\hat{a}|_{[i-1]})) \\
&\leq \sum_{i \in S} (V((a|_S)|_{[i]}) - V((a|_S)|_{[i-1]})) \\
&= V(a|_S) \\
&\leq V(a).
\end{aligned}$$

Here the first inequality is by submodularity, and the second is because V is increasing. So, there is no incentive to lie. □

Proof of Lemma 6. It is not hard to see that $a(q,t)$ is continuous in q. Now, consider any $t > 0$. As noted in the main text, $V(a(q,t)) = t$ for any probability vector q. Let Δ_n denote the probability simplex $\{(q_1, \ldots, q_n) : q_i \geq 0 \text{ and } \sum_i q_i = 1\}$. Define a function $G : A \setminus \{0\} \to \Delta_n$ by rescaling: $G_i(a_1, \ldots, a_n) = a_i/(a_1 + \cdots + a_n)$. Now define $F : \Delta_n \to \Delta_n$ by $F(q) = G(a(q,t))$. This is a continuous map from the simplex Δ_n to itself. Moreover, for each coordinate i, it sends the face $q_i = 0$ of the simplex to itself, since $q_i = 0$ implies $\partial a_i(q,t)/\partial t = 0$ and therefore $a_i(q,t) = 0$ for each t. A result in topology (e.g. Jamison and Ruckle 1976, Lemma 2.1) then implies that F is surjective.

Now to prove the lemma, consider any type $a \neq 0$ (the lemma statement is trivial for $a = 0$) and put $t = V(a)$ in the above. So by surjectivity, there exists some q such that $G(a(q,t)) = F(q) = G(a)$, or equivalently, $a(q,t) = \lambda a$ for some $\lambda > 0$. Moreover, as noted in the main text, $V(a(q,t)) = t$. Thus, combining, we get $V(a) = t = V(a(q,t)) = V(\lambda a)$. However, since V is strictly increasing, this equality can only hold if $\lambda = 1$. Thus we have shown existence of q and t satisfying $a(q,t) = a$, proving the lemma. □

The key to the proof of Proposition 7 is the lemma below:

Lemma A1. *Let $W : [0,\infty)^n \to \mathbb{R}$ be submodular and continuously differentiable. Write W_i for the derivative with respect to coordinate i. Suppose, moreover, that $W_i(t,t,\dots,t) = 0$ for all t and each i.*

Then, for any $t_1,\dots,t_n \in [0,\infty)$ and for each coordinate index i,

$$W(t_i,t_i,\dots,t_i) \le W(t_1,t_2,\dots,t_n).$$

Proof. It suffices to prove the lemma under the assumption that $t_1 \le t_2 \le \cdots \le t_n$; the general statement will then follow by permuting coordinates. Also, it suffices to prove the lemma for $i = n$, and the statement for any other i will follow. This is because $W(t,t,\dots,t)$ is constant as a function of t (since its total derivative with respect to t is $\sum_i W_i(t,t,\dots,t) = 0$).

Define a sequence of n-dimensional vectors by

$$\begin{aligned}
v_1 &= (t_1,t_2,t_3,\dots,t_{n-1},t_n)\\
v_2 &= (t_2,t_2,t_3,\dots,t_{n-1},t_n)\\
v_3 &= (t_3,t_3,t_3,\dots,t_{n-1},t_n)\\
&\vdots\\
v_n &= (t_n,t_n,t_n,\dots,t_n,t_n).
\end{aligned}$$

Now, for each i with $1 \le i < n$,

$$\begin{aligned}
W(v_{i+1}) - W(v_i) &= \int_{t_i}^{t_{i+1}} \left[\frac{d}{dt} W(\underbrace{t,t,\dots,t}_{i}, t_{i+1},\dots,t_n) \right] dt\\
&= \int_{t_i}^{t_{i+1}} \left[\sum_{j=1}^{i} W_j(\underbrace{t,t,\dots,t}_{i}, t_{i+1},\dots,t_n) \right] dt\\
&\le \int_{t_i}^{t_{i+1}} \left[\sum_{j=1}^{i} W_j(t,t,\dots,t) \right] dt\\
&= 0.
\end{aligned}$$

Here, the inequality holds because submodularity implies that each term W_j increases when the k-th argument (for $k > i$) is decreased from t_k to $t \le t_{i+1}$.

Consequently,

$$W(v_n) \le W(v_{n-1}) \le \cdots \le W(v_2) \le W(v_1),$$

which is exactly what we wanted. □

Proof of Proposition 7. Full learning and Bayesian updating are immediate, so we just need to check that incentive compatibility is satisfied. That is, for any probability vector $q = (q_1, \dots, q_n)$, we check that no type a would gain by reporting q instead of following his intended reporting strategy.

Now, for any nonnegative numbers $t_1, \dots, t_n$, write

$$W(t_1, t_2, \dots, t_n) = V(a_1(q, t_1), a_2(q, t_2), \dots, a_n(q, t_n)) - \sum_{i=1}^{n} q_i t_i.$$

This function is submodular in $(t_1, \dots, t_n)$: the $V(\cdots)$ term is a submodular function because it is obtained from the submodular function V by a monotone reparameterization of each coordinate, and the remaining terms are additively separable. Moreover, W is continuously differentiable, with derivatives

$$\frac{\partial W}{\partial t_i} = \left[V_i(a_1(q, t_1), \dots, a_n(q, t_n)) \cdot \left. \frac{\partial a_i}{\partial t} \right|_{(q, t_i)} \right] - q_i.$$

In particular, when all t_i are equal to the same value t, we get

$$\left. \frac{\partial W}{\partial t_i} \right|_{(t,t,\dots,t)} = V_i(a(q, t)) \cdot \left. \frac{\partial a_i}{\partial t} \right|_{(q,t)} - q_i = 0.$$

Hence Lemma A1 applies to W. For each i, apply the lemma, then multiply both sides by q_i, and sum over i. We get

$$\sum_{i=1}^{n} q_i W(t_i, \dots, t_i) \le W(t_1, \dots, t_n). \tag{A1}$$

Noting that $W(t_i, \dots, t_i) = V(a(q, t_i)) - t_i$, and $W(t_1, \dots, t_n) = V(a_1(q, t_1), \dots, a_n(q, t_n)) - \sum_i q_i t_i$, we can add $\sum_i q_i t_i$ to both sides of (A1) to obtain

$$\sum_{i=1}^{n} q_i V(a(q, t_i)) \le V(a_1(q, t_1), \dots, a_n(q, t_n)). \tag{A2}$$

This holds for all $t_1, \dots, t_n$.

Finally, suppose a type a sends message q. Let S be the set of coordinates i such that $q_i > 0$. For each $i \in S$, let t_i be the value such that $a_i(q, t_i) = a_i$ (we observed in the text that such a value exists and is unique). Then, if dimension i is verified, the sender will be believed to be type $a(q, t_i)$, and so will get payoff $V(a(q, t_i))$. For any $i \notin S$, dimension i will not be verified; we may take t_i arbitrary. Then, the left side of (A2) equals the expected payoff that the sender gets by sending message q. Meanwhile, the right side of (A2) equals $V(a|_S) \leq V(a)$. Hence, the deviation gives a payoff of at most $V(a)$, the payoff to following the prescribed strategy. □

Proof of Proposition 8. As a preliminary, we should give the appropriate formulation of the incentive constraint (analogous to (1)) for direct mechanisms in this model. If type z reports a and is tested on dimension i for which $a_i \neq z_i$, trusted verification implies that he necessarily receives a payoff of at least $V(z|_i)$ (and indeed, this bound can be achieved using the belief that places probability 1 on this type). Thus, a verification strategy $p(a)$ can be part of a valid direct mechanism with full learning if and only if

$$V(z) \geq \sum_{i=1}^{n} p_i(a) w_i(a|z), \quad \text{where} \quad w_i(a|z) = \begin{cases} V(a) & \text{if } z_i = a_i \\ V(z|_i) & \text{if } z_i \neq a_i \end{cases} \tag{A3}$$

for all a and z.

Now we proceed to prove existence of the desired direct mechanism. The approach is non-constructive. For each type a, we show that there exist corresponding verification probabilities $p_i(a)$ that satisfy (A3) for all z. By doing this for every a, we form an incentive compatible mechanism.

So fix a type a henceforth. Consider any particular verification probabilities $p = (p_1, \ldots, p_n)$ that sum to 1. Notice that the function

$$U_p(a|z) = \sum_{i=1}^{n} p_i(a) w_i(a|z)$$

is additively separable in the components of z. Therefore, the gain to type z from misreporting as a,

$$G_p(z) = U_p(a|z) - V(z),$$

is supermodular in z.

Notice first that we can reduce the problem to showing existence of p such that $G_p(z) \leq 0$ for all $z \leq a$. Indeed, suppose that this is true, but there is some $x \nleq a$ with $G_p(x) > 0$. Then supermodularity of $G_p(\cdot)$ implies that $G_p(a \wedge x) + G_p(a \vee x) \geq G_p(a) + G_p(x) > 0$, since $G_p(a) = 0$ (here, $\wedge, \vee$ are the componentwise min and max operations). But $a \wedge x \leq a$, which by assertion satisfies $G_p(a \wedge x) \leq 0$; and $G_p(a \vee x) \leq 0$ because $a \vee x \geq a$ implies $V(a \vee x) \geq V(a)$, so type $a \vee x$ cannot gain from the deviation. Contradiction.

Hereinafter, we consider $z \in B = \{z \in A : z \leq a\}$, and use Δ to denote the $(n-1)$-dimensional unit simplex. Suppose, to obtain a contradiction, that for every $p \in \Delta$ there is $z \in B$ such that $G_p(z) > 0$.

For each $p \in \Delta$, let $l_p = \sup_{z \in B} G_p(z)$. We then have $l_p > 0$ for all p, and since $G_p(z)$ is a continuous function of p for any fixed z (moreover, it is Lipschitz continuous with coefficient $V(a)$), l_p is also a continuous function of p. Now, for any p, let

$$D_p = \left\{ z \in B : G_p(z) > \frac{n+1}{n+2} l_p \right\};$$

in other words, D_p is the set of z such that the gain from deviation to a is sufficiently close to the supremum. By definition of l_p, $D_p \neq \varnothing$ for all p.

For each $i \in \{1, \ldots, n\}$, define

$$R_i = \{p \in \Delta : \exists z \in D_p : z_i = a_i\}.$$

Let us show that $R_i \neq \varnothing$ for any i. To do that, we show that $1|_i \in R_i$ (here $1|_i$ means putting probability 1 on component i). Indeed, suppose $1|_i \notin R_i$, then for all $z \in D_{1|_i}$, $z_i < a_i$, and by definition of $G_p(z)$, we have $G_{1|_i}(z) \leq 0$. However, this is impossible for $z \in D_{1|_i}$ by definition of D_p; this contradiction shows that indeed $1|_i \in R_i$.

Introduce the following notation. Let $\|\cdot\|$ denote the sup-norm on $\mathbb{R}^n$, and let $d(x, Y)$ be the distance from point x to nonempty set Y:

$$d(x, Y) = \inf_{y \in Y} \|x - y\|.$$

Now for any $\varepsilon \geq 0$ and nonempty $Y \subset \Delta$, let $N(Y, \varepsilon)$ be the closed ε-neighborhood of set Y, i.e.,

$$N(Y, \varepsilon) = \{p \in \Delta : d(p, Y) \leq \varepsilon\}.$$

Consistently with this definition, $N(Y, 0) = \overline{Y}$, the closure of Y (which equals Y if Y is closed).

Let us now show that $\bigcap_{i=1}^n \overline{R_i} = \varnothing$. Suppose not, then there is some $p \in \bigcap_{i=1}^n \overline{R_i}$. Take $\varepsilon \in \left(0, \frac{1}{n(n+1)} \frac{l_p}{V(a)+1}\right]$ such that for any $r \in N(\{p\}, \varepsilon)$, $l_r \geq \frac{n(n+2)}{(n+1)^2} l_p$; this is possible because l_p is continuous (and the coefficient is smaller than 1). Since $p \in \bigcap_{i=1}^n \overline{R_i}$, for each $i \in \{1, \ldots, n\}$ there is $p^{(i)} \in N(\{p\}, \varepsilon) \cap R_i$. By definition of R_i we can then take $z^{(i)} \in D_{p^{(i)}}$ such that $z_i^{(i)} = a_i$. By definition of $D_{p^{(i)}}$, we have $G_{p^{(i)}}\left(z^{(i)}\right) > \frac{n+1}{n+2} l_{p^{(i)}} \geq \frac{n}{n+1} l_p$. By Lipschitz continuity of $G_p\left(z^{(i)}\right)$ as a function of p (with coefficient $V(a)$), we have

$$\begin{aligned} G_p\left(z^{(i)}\right) &\geq G_{p^{(i)}}\left(z^{(i)}\right) - V\left(a\right)\left\|p - p^{(i)}\right\| \\ &> \frac{n}{n+1}l_p - V\left(a\right)\frac{1}{n\left(n+1\right)}\frac{l_p}{V\left(a\right)+1} \\ &> \left(\frac{n}{n+1} - \frac{1}{n\left(n+1\right)}\right)l_p = \frac{n-1}{n}l_p. \end{aligned}$$

Denote, for any $k \in \{1, \ldots, n\}$, $y^{(k)} = \bigvee_{i=1}^{k} z^{(i)}$; in particular, $y^{(1)} = z^{(1)}$. Let us now show, by induction, that $G_p\left(y^{(k)}\right) > \frac{n-k}{n}l_p$. Indeed, the base case $k = 1$ is already established. Suppose that $G_p\left(y^{(k-1)}\right) > \frac{n-(k-1)}{n}l_p$. Then we have by supermodularity

$$\begin{aligned} G_p\left(y^{(k)}\right) &= G_p\left(y^{(k-1)} \vee z^{(k)}\right) \\ &\geq G_p\left(y^{(k-1)}\right) + G_p\left(z^{(k)}\right) - G_p\left(y^{(k-1)} \wedge z^{(k)}\right) \\ &> \frac{n-(k-1)}{n}l_p + \frac{n-1}{n}l_p - l_p = \frac{n-k}{n}l_p, \end{aligned}$$

where we used $G_p\left(y^{(k-1)} \wedge z^{(k)}\right) \leq l_p$ by definition of l_p. This proves the induction step. Now, taking $k = n$, we have $G_p\left(y^{(n)}\right) > 0$. However, $y^{(n)} = a$, and we get a contradiction, since $G_p\left(a\right) = 0$. This contradiction shows that such p cannot exist, so $\bigcap_{i=1}^{n} \overline{R_i} = \varnothing$.

Now for every $p \in \Delta$, define $E_p = \{q \in \Delta : G_q\left(x\right) \leq 0 \text{ for all } x \in D_p\}$. In other words, E_p is the set of probabilities that make deviation to a unprofitable for all types $x \in D_p$. If we can prove existence of p such that $p \in E_p$ (i.e., a fixed point of mapping $p \mapsto E_p$), then we will reach our desired contradiction, and the proof will be complete. Notice that for every $p \in \Delta$, E_p is closed and convex, because it is the intersection of closed convex sets given by linear inequalities. Also, for every $p \in \Delta$, E_p is nonempty, because $p \notin R_i$ for some R_i (indeed, we showed that $\bigcap_{i=1}^{n} \overline{R_i} = \varnothing$, so $\bigcap_{i=1}^{n} R_i = \varnothing$ as well), in which case vector $1|_i$ is in E_p. If the correspondence E_p were upper-hemicontinuous, we would immediately get existence of a fixed point by Kakutani's theorem. Unfortunately, this might not be true.

Define

$$h = \inf_{p \in \Delta} \max_{i \in \{1,\ldots,n\}} d\left(p, R_i\right);$$

for each p the maximum is finite and well-defined, because each $R_i \neq \varnothing$. Let us show that $h > 0$. Since the infimum is taken over a compact set and the function $d\left(p, R_i\right)$ is continuous in p, it is achieved for some $p \in \Delta$. If $h = 0$, then $d\left(p, R_i\right) = 0$ for all i, and thus $p \in \overline{R_i}$ for all R_i. But we showed that $\bigcap_{i=1}^{n} \overline{R_i} = \varnothing$, which yields a contradiction that proves that $h > 0$. This implies, in particular, that for every point $p \in \Delta$, there is $i \in$

$\{1, \ldots, n\}$ such that within the $\frac{h}{2}$-neighborhood of p there are no points belonging to R_i.

For each $p \in \Delta$, introduce the set Q_p given by:

$$Q_p = \bigcap_{q \in \Delta} N\left(E_q, \frac{2}{h}\|p - q\|\right).$$

We establish the following properties.

First, for every p, $Q_p \subset E_p$, because for $q = p$, $N\left(E_q, \frac{2}{h}\|p - q\|\right) = N(E_p, 0) = \overline{E_p} = E_p$, since E_p is closed.

Second, for every p, Q_p is convex, because it is the intersection of convex sets ($N(Y, \varepsilon)$ is convex for any ε if Y is convex, and E_q is convex for each q).

Third, let $Q \subset \Delta \times \Delta$ be the graph of mapping $p \mapsto Q_p$, i.e.,

$$Q = \{(p, r) \in \Delta \times \Delta : r \in Q_p\};$$

then Q is closed. To see this, notice that

$$\begin{aligned} Q &= \bigcup_{p \in \Delta} \bigcap_{q \in \Delta} \left\{(p, r) : r \in N\left(E_q, \frac{2}{h}\|p - q\|\right)\right\} \\ &= \bigcap_{q \in \Delta} \bigcup_{p \in \Delta} \left\{(p, r) : r \in N\left(E_q, \frac{2}{h}\|p - q\|\right)\right\}. \end{aligned}$$

But for each q, the mapping $p \mapsto N\left(E_q, \frac{2}{h}\|p - q\|\right)$ has a closed graph (this is a continuous set-valued mapping), and thus Q is closed as an intersection of closed sets.

Fourth, for every $p \in \Delta$, Q_p is nonempty. Indeed, from the definition of h it follows that there is $i \in \{1, \ldots, n\}$ such that $q \in N\left(\{p\}, \frac{h}{2}\right)$ implies $q \notin R_i$, and in particular $p \notin R_i$. Let us show that the vector $1|_i$ is in Q_p. To do this, we need to show that for every $q \in \Delta$,

$$1|_i \in N\left(E_q, \frac{2}{h}\|p - q\|\right).$$

If $q \in N\left(\{p\}, \frac{h}{2}\right)$, we have $q \notin R_i$, which implies $1|_i \in E_q$, which establishes the required inclusion for such q. In the complementary case, $q \notin N\left(\{p\}, \frac{h}{2}\right)$, we have $\|p - q\| > \frac{h}{2}$, and thus $N\left(E_q, \frac{2}{h}\|p - q\|\right) = \Delta$ (since E_q is nonempty and the maximum distance between two points in Δ is 1). So the required inclusion is satisfied in this case as well. Since it holds for every q, this proves that $1|_i \in Q_p$, so $Q_p \neq \varnothing$ for any $p \in \Delta$.

Now, the second, third, and fourth properties show that the mapping $p \mapsto Q_p$ satisfies the requirements of Kakutani's fixed-point theorem. Therefore, there is $p \in \Delta$ such that

$p \in Q_p$. The first property now implies that this $p \in E_p$. Therefore, the mapping $p \mapsto E_p$ has a fixed point. We have that for all $x \in D_p$, $G_p(x) \leq 0$, which contradicts the definition of D_p. This contradiction completes the proof. □

Proof of Proposition 9. To show necessity: Suppose such a mechanism exists. Fix a type $a = (a_1, a_2)$. Suppose that when a follows his equilibrium strategy,[12] dimensions 1 and 2 are checked with probability p_1 and p_2 respectively. Now consider any $x_1 < a_1$ and $x_2 < a_2$. If type (a_1, x_2) follows the strategy of type a, with probability p_1 he is believed to be a (due to full learning) and receives payoff $V(a)$; with probability $p_2 = 1 - p_1$ he is believed to be at least $(0, x_2)$ (due to trusted verification). If he instead follows his equilibrium strategy then, by full learning, his payoff is $V(a_1, x_2)$. So incentive compatibility requires

$$p_1 V(a) + (1 - p_1) V(0, x_2) \leq V(a_1, x_2). \tag{A4}$$

Similarly, the incentive of type (x_1, a_2) gives

$$p_1 V(x_1, 0) + (1 - p_1) V(a) \leq V(x_1, a_2). \tag{A5}$$

The first equation implies $p_1 \leq \frac{V(a_1,x_2)-V(0,x_2)}{V(a)-V(0,x_2)}$ (note if the denominator is 0, then by monotonicity $V(a) = V(a_1, x_2) = V(0, x_2)$ and the numerator is also 0). The second likewise implies $p_1 \geq \frac{V(a)-V(x_1,a_2)}{V(a)-V(x_1,0)}$. We thus have $\frac{V(a)-V(x_1,a_2)}{V(a)-V(x_1,0)} \leq \frac{V(a_1,x_2)-V(0,x_2)}{V(a)-V(0,x_2)}$. Cross-multiplying gives

$$(V(a) - V(x_1, a_2))(V(a) - V(0, x_2)) \leq (V(a) - V(x_1, 0))(V(a_1, x_2) - V(0, x_2)),$$

which also holds in either of the zero-denominator cases (since both sides are then zero). Adding $(V(a) - V(x_1, a_2))(V(0, x_2) - V(a_1, x_2))$ to both sides gives the condition in the proposition.

To show sufficiency: Suppose the condition holds for all $x, a \in A$ such that $x < a$. We will construct a direct mechanism that achieves full learning. As argued in the proof of Proposition 8, it suffices to find verification probabilities satisfying (A3).

Fix a, and let us find probability p_1 such that if a report of a leads to verification probabilities $p_1(a) = p_1$, $p_2(a) = 1 - p_1$, this deters all deviations to a. Note that deviation by types x with $x_1 \neq a_1$ and $x_2 \neq a_2$ is automatically deterred, since the deviation will always be detected and the sender will be believed to be either $(x_1, 0)$ or $(0, x_2)$, both of

[12] In fact, one can formulate a revelation principle (analogous to Lemma 0) under the restriction of trusted verification. For brevity, we omit a formal statement.

which are worse than truth-telling. Moreover, types (x_1, a_2) with $x_1 > a_1$ cannot benefit from deviating to a since the truth-telling payoff is $V(x_1, a_2) \geq V(a)$; likewise for types (a_1, x_2) with $x_2 > a_2$. So we need only worry about deviations by types (x_1, a_2) with $x_1 < a_1$ or (a_1, x_2) with $x_2 < a_2$.

Notice that if $V(a) = V(x_1, 0)$ for some $x < a$, then $p_1 = 1$ will work (monotonicity implies $V(a_1, x_2) = V(a)$ for all $x_2 < a_2$, so none of these types gains from deviating, and types (x_1, a_2) will be caught with certainty). Similarly, if $V(a) - V(0, x_2)$ for some $x < a$, then $p_1 = 0$ will work. Thus, we may assume that for any $x < a$, $V(a) > V(x_1, 0)$ and $V(a) > V(0, x_2)$. Again rearranging the terms, the inequality in the proposition statement implies

$$\frac{V(a) - V(x_1, a_2)}{V(a) - V(x_1, 0)} \leq \frac{V(a_1, x_2) - V(0, x_2)}{V(a) - V(0, x_2)}.$$

Since the left-hand side depends on x_1 only and the right-hand side depends on x_2 only, we have

$$\sup_{x_1} \frac{V(a) - V(x_1, a_2)}{V(a) - V(x_1, 0)} \leq \inf_{x_2} \frac{V(a_1, x_2) - V(0, x_2)}{V(a) - V(0, x_2)}.$$

Now if we take $p_1 \in \left[\sup_{x_1} \frac{V(a)-V(x_1,a_2)}{V(a)-V(x_1,0)}, \inf_{x_2} \frac{V(a_1,x_2)-V(0,x_2)}{V(a)-V(0,x_2)}\right]$, we will have that for any x_1 and any x_2,

$$\frac{V(a) - V(x_1, a_2)}{V(a) - V(x_1, 0)} \leq p_1 \leq \frac{V(a_1, x_2) - V(0, x_2)}{V(a) - V(0, x_2)}.$$

Rearranging brings us back to conditions (A4)–(A5), which coincide with the incentive constraints (A3) for types (a_1, x_2) and (x_1, a_2). So deviations to a by these types are deterred. □

Proof of Proposition 10. We just need to check that if condition (1) is satisfied for the function V, then it is also satisfied for $V' = U \circ V$. For any $a, \hat{a}$, put $\lambda = \sum_{i:\hat{a}_i = a_i} p_i(\hat{a})$; thus $\lambda \in [0, 1]$. Condition (1) says that $V(a) \geq \lambda V(\hat{a})$. Then,

$$U(V(a)) \geq U(\lambda V(\hat{a})) \geq \lambda U(V(\hat{a}))$$

where the first inequality is because U is increasing and the second is because U is concave (and must map 0 to 0 in order for $U \circ V$ to be an allowable payoff function). Thus, (1) holds for $U \circ V$. □

Proof of Proposition 11. Let $\tilde{V}(a) = V(a; a)$. By Proposition 2, there is a valid direct mechanism that achieves full learning for payoff function $\tilde{V}(a)$; denote such a mechanism by $\mathcal{M}$. Let us show that this same mechanism would remain incentive compatible if the payoffs of type x if he is believed to be type a were given by $V(a; x)$.

Suppose not, so that some type x prefers to deviate and report type $a \neq x$. This immediately implies $V(a;x) > V(x;x)$, for otherwise this deviation would not have any upside. Since we assumed that $V(a;x) \leq \max\{V(a;a), V(x;x)\}$, it must be that $V(a;x) \leq V(a;a)$. Given that $\mathcal{M}$ is a valid mechanism for the payoff function $\tilde{V}(a)$, it must be that

$$\tilde{V}(x) \geq \left(\sum_{i:\ a_i = x_i} p_i(a)\right) \tilde{V}(a) + \left(\sum_{i:\ a_i \neq x_i} p_i(a)\right) \tilde{V}(0)$$

(the last term $\tilde{V}(0)$ is zero, but we wrote it out explicitly). Since $V(a;x) \leq V(a;a) = \tilde{V}(a)$ and $V(0;x) = \tilde{V}(0)$, this implies

$$V(x;x) \geq \left(\sum_{i:\ a_i = x_i} p_i(a)\right) V(a;x) + \left(\sum_{i:\ a_i \neq x_i} p_i(a)\right) V(0;x).$$

In other words, the deviation to reporting a is not profitable. This proves that $\mathcal{M}$ is a valid mechanism. □

Finally, we formalize the claim in Footnote 7, that full learning of $V(a)$ implies full learning of a is possible. For this we must return to the original formulation of the model, where posterior beliefs are non-degenerate, and V is defined on $\Delta(A)$. We need an extra assumption: Say that V *respects constant values* if, for every constant c, if μ is any distribution on A such that $V(a) = c$ for all a in the support of μ, then $V(\mu) = c$ as well.

Say that an (indirect) mechanism achieves *full learning of* $V(a)$ if, for every type a, every history $h \in H(a|\mathcal{M})$, and every $a' \in \text{supp}(\mu(h))$, we have $V(a') = V(a)$.

Proposition A2. *Assume that V respects constant values. If there exists an indirect mechanism that achieves full learning of $V(a)$, then there exists a direct mechanism with full learning of a.*

Proof. Let $\mathcal{M} = (M, \sigma, p, \mu)$ be the mechanism that achieves full learning of $V(a)$. We now repeat the proof of Lemma 0. The same proof goes through, except for two adjustments: in the step that originally applied full learning for the original mechanism, we now apply full learning of $V(a)$ together with respecting constant values; and the fact that type a receives equilibrium payoff $V(a)$ in the original mechanism also uses these two properties. □