Theoretical Economics 12 (2017), 175–210 1555-7561/20170175

Directives, expressives, and motivation

Toru Suzuki
Economics Discipline Group, University of Technology Sydney

When an agent's motivation is sensitive to how his supervisor thinks about the agent's competence, the supervisor has to take into account both informational and expressive contents of her message to the agent. This paper shows that the supervisor can credibly express her trust in the agent's ability only by being unclear about what to do. Suggesting what to do, i.e., "directives," could reveal the supervisor's "distrust" and reduce the agent's equilibrium effort level even though it provides useful information about the decision environment. There is also an equilibrium in which directives are neutral in expressive content. However, it is shown that neologism proofness favors equilibria in which directives are double-edged swords.

Keywords. Communication games, directives, expressives, economics and language.

JEL classification. D83.

1. Introduction

Consider a manager who tells a worker what to do for every single task. Even if the manager's message can resolve uncertainty about the decision environment, such a message could have an adverse effect on the worker's motivation when it sounds disrespectful. Words could convey the speaker's attitude toward the listener, i.e., expressive content, and affect the listener's decision in practice. In fact, the importance of such a psychological effect of words has been known for a long time: in Aristotle's *Rhetoric*, such an effect—*pathos*—is treated as one of the most important forces of words. The purpose of this paper is to analyze the trade-off between the informational and expressive contents of messages in a simple communication game.

This paper focuses on a communication problem between a supervisor (she) and an agent (he) that is applicable to manager–worker, teacher–student, and parent–children communications. The agent has two decisions to make: choosing (i) what to do and (ii) how much effort to put in. The agent is most productive when what he does matches the state of nature. Both the supervisor and the agent prefer higher output while the

Toru Suzuki: toru.suzuki@uts.edu.au
I am grateful to two anonymous referees and the co-editor for insightful suggestions. I also thank Francoise Forges, Piero Gottardi, Susumu Imai, Claudio Mezzetti, Masahiro Okuno-Fujiwara, Jonathan Newton, Masatoshi Tsumagari, and participants of seminars in Queen's University in Belfast, Keio University, UNSW, University of Queensland, University of Sydney, and UTS for helpful discussions and comments. I owe special thanks to Bart Lipman and Joel Sobel for detailed comments.

DOI: 10.3982/TE1843

agent also takes into account the disutility of effort. The supervisor who does not know whether the agent knows the state sends a costless message to the agent.

Note that there is no conflict of interest in communicating about the state. Thus, with the standard assumption that payoffs depend only on the action and the state, the only relevant thing to communicate is the state. As a result, the supervisor communicates efficiently about the state in the most natural equilibrium of the game, that is, she suggests what to do. However, the supervisor's problem becomes nontrivial if the expressive content of the message could affect the agent's action. Specifically, suppose that making additional effort becomes more painful for the agent if he feels the supervisor's distrust in his ability. The supervisor then has to take into account how the agent could interpret the expressive content of her message in equilibrium.

I analyze the communication problem as a cheap talk game in two dimensions. The agent is either competent or incompetent depending on whether he can observe the state of nature. The supervisor who does not know the agent's competence sends a message given the state and her belief about the agent's competence. The agent who does not know the supervisor's belief about his competence updates his belief about the supervisor's belief given her message. If the agent is incompetent, he also updates his belief about the state given her message. The agent then makes decisions based on his updated belief.

To analyze the game, this paper employs two useful concepts, *directives* and *expressives* in philosophy of language. A directive message reveals the true state and induces the right action; an expressive message expresses the supervisor's belief about the agent and could affect the agent's "motivation." In this paper, whether a message is directive (expressive) is determined in equilibrium; this approach allows a message to be both directive and expressive.

The game has an equilibrium in which the supervisor always uses a directive message as in the standard setting. However, since the supervisor has an incentive to pretend that she trusts the agent's ability, a question is whether she can credibly express her belief about the agent's competence in equilibrium. It is shown that such an "expressive equilibrium" always has a handicapping property; that is, the supervisor who trusts the agent's ability always communicates inefficiently about the state to express her trust. In other words, whenever the supervisor credibly communicates her belief about the agent's ability, she needs to sacrifice informativeness about the state.

Since the supervisor has various ways to handicap herself in the two-dimensional communication, many kinds of expressive equilibria can exist. To provide further insights, I focus on expressive equilibria with some degree of consistency in expressive content; that is, whether the expressive content of a message is more positive than that of another message does not depend on the receiver's competence. Intuitively, in these equilibria, when a message means trust for the competent agent, the message does not mean distrust for the incompetent agent. This consistency property is appealing if we consider the use of a natural language; when the supervisor uses messages so that the equilibrium meaning respects the preexisting meaning to some degree, the equilibrium possesses the consistency property. It is shown that any directive message has negative expressive content in any expressive equilibrium with the consistency property. The

result might explain why directives could sound offensive in ordinary communication even though more information about the decision environment should not be harmful to the decision maker.

Since there is no conflict of interest in communicating about the state, the main purpose of inefficient communication seems to be to express trust. Thus, I consider expressive equilibria in which whenever the supervisor communicates about the state inefficiently, the purpose is purely expressive; that is, the use of the messages does not depend on the state. To characterize such state-independent expressive equilibria, I introduce an intuitive communication strategy called *simple expressive*: the supervisor uses directives when her trust in the agent's ability is lower than a certain level while her message is uninformative about the state if the level of her trust is higher. I show that any state-independent expressive equilibrium is payoff-equivalent to a simple expressive equilibrium. I also provide a sufficient condition for the existence of a simple expressive equilibrium; a simple expressive equilibrium exists whenever the agent's choice of what to do is sufficiently important and difficult. Intuitively, the condition guarantees that the supervisor can effectively handicap herself to make her expressive message credible.

While directives have negative expressive content in natural expressive equilibria, there also exists an equilibrium that consists entirely of "pure directives," which have no expressive content. However, introspection suggests that the equilibrium with pure directives might not be so stable. Suppose that the supervisor says "I trust you" in an organization where the use of pure directives is their convention. If this message could effectively handicap herself by inducing an suboptimal choice, it might be able to convey her trust overriding the equilibrium. To investigate this observation formally, I employ the concept of neologism-proof equilibrium. It is shown that whenever a simple expressive equilibrium exists, the equilibrium that consists entirely of pure directives is not neologism-proof. Moreover, simple expressive equilibria are neologism-proof whenever the agent's payoff function is not too sensitive to the supervisor's trust in the agent's ability. Thus, neologism proofness favors simple expressive equilibria over those with pure directives.

Related literature. This paper contributes to the literature on economics and language, especially pragmatics in economic contexts.[1] Rubinstein (2000) and Glazer and Rubinstein (2001, 2006) analyzed a persuasion problem by extending the idea of Grice (1989) to a noncooperative situation. In their models, the speaker's statement is verifiable, and the listener chooses a *persuasion rule* that specifies how the listener responds to each statement. The optimal persuasion rule is then determined as the solution of the listener's optimization problem. Since the optimal persuasion rule implicitly captures how the listener interprets each statement, their approach essentially pins down the meaning of statements in the noncooperative context. The current paper, on the other hand, analyzes an advising problem in which the supervisor has to manage not only the informational but also the expressive content of a message to maximize the agent's output. Since the supervisor's trust is private information and unverifiable, my

[1] Pragmatics is the branch of linguistics that analyzes uses of language.

model belongs to a class of cheap talk games introduced by Crawford and Sobel (1982), and the main interest is in the trade-off between the informational and expressive contents of a message. Thus, the relevant concept in pragmatics is different from Glazer and Rubinstein's.

This paper borrows two useful concepts—directives and expressives—from pragmatics. These concepts were originally introduced by Searle (1975) to classify effects of utterances a speaker intends, i.e., "illocutionary acts."[2] A directive message is an utterance that intends to suggest or order a certain action; an expressive message is an utterance that intends to express the speaker's feeling or attitude. Importantly, the speaker's intended meaning of a sentence does not always coincide with the literal meaning but it is determined by the use of the sentence given a context. Thus, the current paper defines directives and expressives as properties of a communication strategy so that whether a message is directive or/and expressive is determined in equilibrium. The distinction between directives and expressives is useful since their roles are different in this model: a directive message resolves uncertainty of decision environments while an expressive message communicates about the supervisor's attitude and affects the agent's motivation.

Another topic that is closely related to my paper is politeness. One of the major approaches to understanding politeness in communication is based on "face-saving" by Brown and Levinson (1987). According to the theory, some politeness is intended to avoid "face-threatening acts." For example, a sensible speaker knows that a directive message could threaten a listener's face and so avoids it or uses an indirect expression. If the agent's belief about how the supervisor views the agent's competence is a proxy for "face," the current paper can be interpreted as a communication game with face-saving. Unlike the original approach, which relies on introspection to identify face-threatening acts, my approach explains the conditions under which directives could be face-threatening in equilibrium.

The idea that the sender's private information interferes with "factual language" is similar to Blume and Board (2013). In their model, the sender's set of available messages depends on his language type. Since the sender's language type is private information, and his message reflects not only the state but also his language type, the sender cannot communicate about the state efficiently even in the common interest game. In the current paper, the sender's private information is her belief about the receiver's competence. Thus, the fact that the supervisor uses a directive message could reveal her distrust in the agent's competence. The difference between the current paper and their paper is not only in the setting but also in the nature of the result. Efficient communication about the state is precluded in their model whereas it is an equilibrium outcome that can be susceptible to an intuitive neologism in the current paper. Thus, whether factual language emerges as a reasonable prediction depends on the situation in the current paper. There are also other cheap talk models in which a potentially biased sender tries to communicate about the state, e.g., Morris (2001) and Morgan and Stocken (2003). However,

[2]The term "illocutionary act" was introduced by Austin (1962). Searle (1975) categorized illocutionary acts into five classes: directives, expressives, commissives, assertives, and declaratives.

as in Blume and Board (2013), these models preclude efficient communication about the state.

The basic idea that cheap talk can credibly signal one aspect by sacrificing the other aspect is similar to Chakraborty and Harbaugh (2010). Unlike in their model, the sender does not "talk down" one aspect to credibly signal the other aspect. Instead, the sender makes her message credible by being uninformative about the other aspect in the current paper; that is, the opacity of a message expresses the supervisor's trust.

In the current paper, the sender can credibly express her trust since she does not know the receiver's type. The idea that multiple receivers make communication informative is similar to the notion of "mutual discipline" in Farrell and Gibbons (1989). The difference is that unlike in their model, the two possible receivers in the current paper have the identical payoff function but they are different only in private information.

Finally, this paper also contributes to the literature that studies incentive problems beyond extrinsic motivation. Bénabou and Tirole (2003) analyze how an incentive scheme should be designed when higher extrinsic incentive could damage the worker's intrinsic motivation. This paper, on the other hand, analyzes how the supervisor can manage the worker's motivation with cheap talk. In my model, a directive message could be a double-edged sword: it provides useful information about the decision environment while it could also damage the worker's motivation because of the expressive content.

2. Model

2.1 *Basics*

There is a supervisor (she) and an agent (he). Let Ω be a finite set of states of nature and let $\pi(\omega)$ be a common prior of $\omega \in \Omega$. The supervisor knows ω whereas the agent may or may not know ω. Specifically, the agent is either competent or incompetent: the agent is competent if he knows ω whereas he is incompetent if he does not know ω. The agent has two decision problems. The first problem is technology choice. In this problem, he chooses technology a from $A = \{a_\omega\}_{\omega\in\Omega} \cup \{a_0\}$, where a_ω is the optimal technology in state ω, and a_0 is a safe technology. The agent's second problem is to choose his effort level e from $E = [\underline{e}, \bar{e}] \subset \Re_+$ given his technology choice a.

The agent's output depends on (i) the match between technology a and state ω and (ii) his effort level e. If the agent chooses the right technology given ω, i.e., $a = a_\omega$ at ω, then the output given e is $y(e)$, where $y : E \to \Re_+$. It is assumed that $y(e)$ is twice differentiable, $y'(e) > 0$ and $y''(e) \leq 0$ for any $e \in E$. If the agent chooses a wrong technology, i.e., $a = a_{\omega'}$ at $\omega \neq \omega'$, his productivity is discounted by $\delta \in [0, 1)$. That is, the output given e is $\delta y(e)$. Finally, if the agent chooses the safe technology, i.e., $a = a_0$, the productivity is discounted by $\lambda \in (0, 1)$, and the output given e is $\lambda y(e)$ independently of ω. It is assumed that $\lambda > \delta$ so that the productivity is lowest under wrong choices. Let $f(e, a, \omega)$ be the output function that summarizes the above specification.

The supervisor does not know whether the agent is competent. Let p be the supervisor's prior belief that the agent is competent. That is, p measures the supervisor's "trust" in the agent's competence. The agent does not know how the supervisor thinks

about his competence. In other words, p is the supervisor's private information. Assume that p follows distribution $\Psi(p)$ with continuous density $\psi(p)$ and $\text{supp}(\psi) = P = [0, 1]$. Furthermore, p and ω are independent, and Ψ is common knowledge.

The agent's disutility of effort depends on the supervisor's trust in the agent's competence. Specifically, an additional effort gives him higher disutility if he feels that he is less likely to be trusted. This might be because it is less enjoyable to work under the supervisor who does not trust the agent's competence.[3] Formally, let $c(e, p)$ be the agent's disutility function, which is twice differentiable in each argument.[4] The marginal disutility of effort is strictly positive and increasing in e given any p, i.e., $c_1(e, p) > 0$ and $c_{11}(e, p) > 0$ for any (e, p). Moreover, so as to capture the negative effect of distrust on the agent's motivation, assume that $c_{12}(e, p) < 0$ for any (e, p). Finally, assume $c_1(\bar{e}, 1) > y'(\bar{e})$ and $c_1(\underline{e}, 0) < y'(\underline{e})$ so that the optimal effort level of the competent agent is in the interior of E given any belief about p.

The agent's output is shared by both players, and the supervisor prefers a higher output level independently of (p, ω). Specifically, the supervisor's payoff given (e, a, ω) is $f(e, a, \omega)$. Since the effort is costly for the agent, the agent's payoff given (e, a, ω, p) is $f(e, a, \omega) - c(e, p)$.[5]

The game proceeds as follows: the supervisor observes (p, ω) and sends message m from a sufficiently rich set M. Then, given the message and his competence, the agent chooses technology $a \in A$ and his effort level $e \in E$.

Remark 1. To capture the effect of a message on the agent's motivation, this paper assumes that the agent intrinsically cares about the supervisor's belief as in Bernheim (1994), Köszegi (2006), and other psychological games.[6] Note that a message does not affect the agent's disutility directly. Instead, it affects the agent's payoff through his equilibrium belief about the supervisor's opinion. This approach allows us to analyze how the expressive content of a message depends on the parameters that capture the physical setting behind the communication.

The belief-dependent payoff can be also interpreted as a reduced form of a dynamic interaction with the standard payoff function.[7] However, since the expressive content of

[3] In social psychology, Enzle and Anderson (1993) found that surveillance could undermine intrinsic motivation. Moreover, the effect was more pronounced when the subjects interpreted the surveillance as a stronger signal of distrust.

[4] Since the agent does not observe p, $c(e, p)$ is the agent's disutility of e when he knows p hypothetically. Thus, the agent's decision making is based on his expected disutility.

[5] The output is implicitly treated as a nonrivalrous benefit to the supervisor and the agent. However, the results of this paper are preserved even if the output is a profit, which is allocated to each player according to a sharing rule.

[6] In psychological games (Geanakoplos et al. (1989)), payoffs depend on "action beliefs." As in Bernheim (1994), payoffs depend on "type beliefs" in the current paper. Battigalli and Dufwenberg (2009) introduce dynamic psychological games that can incorporate type-belief-dependent payoffs.

[7] From the perspective of evolutionary psychology, Nature might give us a belief-dependent preference to deal with complicated social interaction problems under a cognitive constraint.

words could affect the listener's decision even without any future interaction in practice, the psychological interpretation might be more natural.

2.2 *Strategy and equilibrium*

The supervisor sends $m \in M$ given (p, ω). The supervisor's communication strategy is then a mapping:[8]

$$\sigma : P \times \Omega \to M.$$

The agent observes private signal h: if he is competent, he observes the true state, i.e., $h = \omega$, whereas $h = n$ if he is incompetent. Let $H = \Omega \cup \{n\}$. The agent's technology choice function is

$$a : M \times H \to A.$$

Moreover, the agent's effort choice function is

$$e : M \times H \to E.$$

I employ *perfect Bayesian equilibrium* to analyze this game. Specifically, first, let $e(m, h|a)$ be the agent's optimal effort level given (m, h) when his technology is a. That is, $e(m, h|a)$ solves

$$\max_{e \in E} \left\{ \sum_{\omega} f(e, a, \omega) \mu_{\omega}(\omega|m, h) - \int_{p} c(e, p) \mu_{p}(p|m, h)\, dp \right\},$$

where $\mu_{\omega}(\omega|m, h)$ and $\mu_{p}(p|m, h)$ are the agent's posterior beliefs about ω and p, respectively. From the assumptions about $y(e)$ and $c(e, p)$, the expected marginal output is decreasing in e whereas the expected marginal disutility is strictly increasing in e given any μ_{ω} and μ_{p}. Thus, for any (m, h), $e(m, h|a)$ is unique given a.

Given $\{e(m, h|a)\}_{a \in A}$, the agent solves the problem

$$\max_{a \in A} \left\{ \sum_{\omega} f(e(m, h|a), a, \omega) \mu_{\omega}(\omega|m, h) - \int_{p} c(e(m, h|a), p) \mu_{p}(p|m, h)\, dp \right\}.$$

Since the competent agent knows the true state, his optimal technology choice is always a_{ω} at ω. On the contrary, the incompetent agent could have more than one optimal technology depending on μ_{ω}. Let $a^*(m, h)$ be the agent's optimal technology choice function that specifies his choice given (m, h). Moreover, let $e^*(m, h) = e(m, h|a^*(m, h))$ be the optimal effort level given (m, h).[9]

The supervisor with (p, ω) believes that the agent is competent with probability p. Thus, her expected payoff from m given strategy σ is

$$U^m(p, \omega) = p f(e^*(m, \omega), a^*(m, \omega), \omega) + (1 - p) f(e^*(m, n), a^*(m, n), \omega).$$

[8] Since allowing mixed strategies does not add any insight into the results of this paper, we focus on pure strategies.

[9] Even when the incompetent agent has more than one optimal technology, the optimal effort level is the same under any optimal technology.

Then $\sigma^*(p, \omega)$ is an equilibrium communication strategy if, for any $(p, \omega) \in P \times \Omega$,

$$U^{\sigma^*(p,\omega)}(p, \omega) \geq U^m(p, \omega)$$

for all $m \in M$.

REMARK 2. Since the competent agent knows ω, some "off-equilibrium message" for the competent agent can be an "equilibrium message" for the incompetent agent. To see this, suppose $\sigma(p', \omega') = m'$ but $\sigma(p, \omega'') \neq m'$ for any p. If the competent agent receives m' at ω'', he cannot apply Bayes' rule since there is no p that sends m' at ω''. On the contrary, if the incompetent agent receives m', he can still apply Bayes' rule since the message could be sent by the supervisor with (p', ω').

3. EQUILIBRIUM ANALYSIS

Since the supervisor's payoff function does not directly depend on m, there always exist equilibria in which the supervisor's message conveys no information about (p, ω). This section analyzes equilibria that reveal some information about (p, ω).

3.1 *Fully directive equilibrium*

With payoff functions that do not depend on the other player's belief, there is no conflict of interest in the communication. Consequently, the only relevant information for both players is ω, and the supervisor reveals ω in the most natural equilibrium. Since the supervisor knows that revealing ω always induces the right technology choice, this is essentially a directive message. To formalize the notion of directives, let

$$P_\sigma(m, \omega) = \{p \in P : \sigma(p, \omega) = m\}.$$

Given communication strategy σ, $m' \in M$ is *directive* if there exists ω' such that $P_\sigma(m', \omega') \neq \varnothing$ and $P_\sigma(m', \omega) = \varnothing$ for any $\omega \neq \omega'$. In short, a message is directive if the agent can uniquely identify the state when he receives it.

The simplest communication strategy with directives is making a suggestion at every state independently of p. Formally, a communication strategy is *fully directive* if, for each ω, there exists message $m_\omega \in M$ such that $P_\sigma(m_\omega, \omega) = P$ and $P_\sigma(m_\omega, \omega') = \varnothing$ for any $\omega' \neq \omega$. Moreover, an equilibrium is a *fully directive equilibrium* if the supervisor uses a fully directive strategy. In this equilibrium, each message perfectly reveals state ω while it is uninformative about p; that is, the directives have no expressive content.

The first result states that the most natural equilibrium under "belief-independent payoffs" is preserved in this model.

FACT 1. *There exists a fully directive equilibrium.*

In this equilibrium, the supervisor's payoff is always $y(\tilde{e})$, where

$$\tilde{e} = \arg\max_{e \in E} \left\{ y(e) - \int_p c(e, p)\psi(p)\, dp \right\}.$$

In short, $\tilde{e}$ is the effort level induced by directives without expressive content.

3.2 *Expressive equilibrium*

The supervisor always wants the agent to believe that p is high while her payoff function does not depend on m. Thus, it is not obvious whether the supervisor can credibly express her trust in the agent's ability. Given a communication strategy σ, $m' \in M$ is *expressive* if $P_\sigma(m', \omega) \subsetneq P$ and $P_\sigma(m', \omega) \neq \varnothing$ for some ω. In short, an expressive message reveals some information about p. An equilibrium is an *expressive equilibrium* if some (p, ω) sends an expressive message. Since this is a cheap talk game, there can be an expressive equilibrium in which no message can influence the agent's effort level. However, such an equilibrium is not interesting as it is payoff-equivalent to some equilibrium that is pooling in p. Thus, our interest is in expressive equilibria in which the expressive content of some message influences the agent's effort. An expressive equilibrium is *influential* if there exist $p, p' \in P$ such that $e(\sigma(p, \omega), \omega) \neq e(\sigma(p', \omega), \omega)$ for some ω.

PROPOSITION 1. *In any influential expressive equilibrium, there exists* $(p', \omega') \in P \times \Omega$ *such that* $\sigma(p, \omega')$ *is not directive for all* $p \in (p', 1]$.

Most proofs are given in the Appendix.

The proposition states that any influential expressive equilibrium involves some inefficient communication about the state. The idea is that since the supervisor has an incentive to exaggerate her trust, she can credibly express her trust only by sacrificing informativeness about the state. In other words, the supervisor needs to handicap herself by communicating inefficiently about the state to express her trust.

The supervisor has various ways to handicap herself in this multidimensional communication. As a result, the use of expressive messages could be complex in some expressive equilibria; for example, whether the expressive content of a message is positive or negative could depend on the receiver's competence. Since such a lack of consistency in expressive content could make the equilibrium difficult to play, one of the properties of natural expressive equilibria seems to be some degree of consistency in expressive content. Specifically, if the expressive content of m' is more positive than that of m'' for the competent agent, the incompetent agent also finds m' more positive.[10]

To formalize the property, let $e^c(m, n)$ be the solution of the problem

$$\max_{e \in E} \left\{ y(e) - \int c(e, p) \mu_p(p|m, n)\, dp \right\}.$$

That is, this is the optimal effort level when the agent knows ω but his belief about p conditional on m is the same as the incompetent agent's. In short, $e^c(m, n)$ extracts the effect of the expressive content of m on the incompetent agent's effort choice. An influential expressive equilibrium is *consistent* if, for any equilibrium messages m' and m'' such that $e(m', \omega) \geq e(m'', \omega)$ for some ω, $e^c(m', n) \geq e^c(m'', n)$.

[10] This property is appealing when we consider the use of a natural language: when messages have preexisting meanings, the equilibrium meaning of each message would respect the preexisting meaning to some degree in the focal equilibrium.

LEMMA 1. *In any consistent expressive equilibrium, if $\sigma(p', \omega)$ is not directive, then $\sigma(p, \omega)$ is not directive for all $p > p'$.*

To provide intuition, observe that, in any consistent expressive equilibrium, the competent agent's effort level is strictly higher than the incompetent agent's whenever the supervisor sends an uninformative message. Then, since the supervisor with a low (high) p believes that the agent is competent with a low (high) probability, the supervisor's expected payoff from an uninformative message is increasing in p. On the contrary, note that the supervisor's payoff from a directive message is constant in p. Hence, the supervisor prefers to send an uninformative message whenever a lower p sends an uninformative message.

From Lemma 1, if the supervisor with p' at ω sends a directive message in a consistent expressive equilibrium, the supervisor with $p < p'$ at ω also sends a directive message. Hence, if the supervisor sends a directive message and a "nondirective" message at ω, the expressive content of the directive message is always negative relative to that of the nondirective message in the consistent expressive equilibrium.

COROLLARY 1. *In any consistent expressive equilibrium, if $\sigma(p_1, \omega)$ is directive while $\sigma(p_2, \omega)$ is not directive, then $e(\sigma(p_1, \omega), \omega) < e(\sigma(p_2, \omega), \omega)$.*

Since there is no conflict of interest in communicating about the state, the only motivation to use a nondirective message seems to be to express her trust. In fact, in ordinary communication, an expressive sentence such as "I trust you" is often used as a "pure expressive," that is, the message does not convey any information about the decision environment. Let

$$\Omega_\sigma(m) = \{\omega : \sigma(p, \omega) = m \text{ for some } p\}.$$

An influential expressive equilibrium is *state independent* if, for any $m \in M$ such that $|\Omega_\sigma(m)| > 1$, $P_\sigma(m, \omega) = P_\sigma(m, \omega')$ for any $\omega, \omega' \in \Omega$. That is, in a state-independent expressive equilibrium, any nondirective message does not reveal any information about ω. Clearly, any state-independent expressive equilibrium is a consistent expressive equilibrium.

The next lemma provides the property of state-independent expressive equilibrium: the supervisor never uses a message that can make the incompetent agent choose a wrong technology.

LEMMA 2. *In any state-independent expressive equilibrium, there is no $(p, \omega) \in P \times \Omega$ such that $a(\sigma(p, \omega), n) = a_{\omega'} \neq a_\omega$.*

From Proposition 1, the supervisor makes her expressive message credible by handicapping herself. If the supervisor induces the incompetent agent to choose $a_{\omega''}$ at ω', the same message induces the right technology at ω''. Then, since sending the message at ω' is more "costly" than sending it at ω'', the expressive content of the message should depend on the state. Thus, in any state-independent expressive equilibrium, the supervisor never uses a message that can induce a wrong choice.

Now I introduce an intuitive communication strategy in which the uninformative message is state independent. In this strategy, the supervisor with a low p uses directives while the supervisor with a high p sends a message that is uninformative about ω. Formally, a communication strategy is *simple expressive* if

$$\sigma(p,\omega)=\begin{cases} m_{\varnothing} & \text{if } p\in(\hat{p},1] \\ m(\omega) & \text{if } p\in[0,\hat{p}), \end{cases}$$

where $m(\omega)$ is an injection from Ω to $M\backslash\{m_{\varnothing}\}$. In short, $m(\omega)$ is a directive message; that is, if the agent receives $m(\omega')$, he knows that the state is ω'. An equilibrium is a *simple expressive equilibrium* if the supervisor uses a simple expressive strategy. In this equilibrium, the uninformative message $m_{\varnothing}$ expresses the supervisor's trust in the agent's competence by being completely uninformative about ω. In ordinary communication, it can be interpreted as an expressive sentence whose literal meaning does not refer to any state, e.g., "I trust your ability." On the contrary, all the directive messages have negative express content; that is, they reveal that p is lower than $\hat{p}$.

PROPOSITION 2. *Any state-independent expressive equilibrium is payoff-equivalent to a simple expressive equilibrium.*

The logic behind this result is as follows. From Lemma 2, if the supervisor uses a message at more than one state, the message has to make the incompetent agent choose the safe technology. Moreover, from the state-independence property, the expressive message has to be used for all states. Then, from Lemma 1, there exists $\hat{p}$ such that the supervisor with $p<\hat{p}$ sends some directive message at each state while the supervisor with $p>\hat{p}$ sends some expressive message that makes the incompetent agent choose the safe technology. When the supervisor uses a communication strategy that is convex in p, this is a simple expressive equilibrium with cutoff $\hat{p}$. It can be shown that even if the supervisor uses a strategy that is nonconvex in p, it is payoff-equivalent to the simple expressive equilibrium with $\hat{p}$.

The existence of a simple expressive equilibrium is not guaranteed. To provide a sufficient condition for the existence of a simple expressive equilibrium, let

$$e_0=\arg\max_{e\in E}\{y(e)-c(e,0)\}$$

and

$$e_\lambda=\arg\max_{e\in E}\left\{\lambda y(e)-\int_p c(e,p)\psi(p)\,dp\right\}.$$

That is, e_0 is the optimal effort level under the right technology when the agent is completely distrusted. Moreover, e_λ is the optimal effort level under the safe technology when the agent has no information about how the supervisor thinks about him. Let $\omega_{\max}\in\arg\max_\omega \pi(\omega)$.

PROPOSITION 3. *There exists a simple expressive equilibrium if* $(1-\delta)\pi(\omega_{\max})+\delta\le\lambda< y(e_0)/y(e_\lambda)$. *Furthermore, if* $(1-\delta)\pi(\omega_{\max})+\delta>\lambda$, *then there is no simple expressive equilibrium.*

Intuitively, Proposition 3 states that a simple expressive equilibrium exists if the technology choice problem is sufficiently "difficult" and "important." First, if $(1-\delta)\pi(\omega_{\max})+\delta \leq \lambda$, $m_\varnothing$ induces the incompetent agent to choose the safe technology. This condition is satisfied when the technology choice problem is sufficiently difficult given δ and λ. To see this, suppose $|\Omega|=5$. If $\pi(\omega_{\max})=0.2$, the technology choice problem is difficult for the incompetent agent since all states are equally likely to happen. On the contrary, when $\pi(\omega_{\max})=0.9$, the choice problem is easier since the true state is very likely to be $\omega_{\max}$. Second, if $\lambda < y(e_0)/y(e_\lambda)$, managing the incompetent agent's technology choice is more important than managing the expressive content of directives. To see this, recall that $y(e_0)$ is the output level when the supervisor uses a directive message and the agent thinks that he is completely distrusted. Moreover, $\lambda y(e_\lambda)$ is the output level when the incompetent agent chooses the safe technology without any information. Thus, if $\lambda < y(e_0)/y(e_\lambda)$, the loss from inducing the safe technology instead of the right technology is larger than the loss from being perceived as the worst type, i.e., $p=0$.

The equilibrium cutoff type is constructed so that the cutoff type's payoffs from m_ω and $m_\varnothing$ are indifferent. Since the cutoff type's payoffs from m_ω and $m_\varnothing$ are increasing in the cutoff, the uniqueness of the equilibrium cutoff is not guaranteed. However, it can be shown that if the disutility function is not so sensitive to p, the equilibrium cutoff is unique and $\lambda < y(e_0)/y(e_\lambda)$ becomes a necessary condition for the existence of a simple expressive equilibrium.

The equilibrium effort levels in simple expressive equilibrium depend on the productivity of the safe technology, i.e., λ. To provide the comparative statics, suppose that there exists a unique simple expressive equilibrium outcome. Let $e(m,h;\lambda)$ be the effort level in the simple expressive equilibrium conditional on (m,h) under λ.

Fact 2. *Suppose $\lambda'' > \lambda'$ and that there exists a unique equilibrium cutoff under λ' and λ''. Then $e(m_\varnothing,\omega;\lambda') > e(m_\varnothing,\omega;\lambda'')$ and $e(m_\omega,h;\lambda') > e(m_\omega,h;\lambda'')$ for $h=n,\omega$.*

Fact 2 implies that the expressive content of $m_\varnothing$ and m_ω is more positive under λ'. To see this, note that when the safe technology becomes less productive, the supervisor's expected payoff from $m_\varnothing$ gets lower whereas it does not affect her expected payoff from m_ω. Thus, the lower λ encourages more p to send m_ω instead of $m_\varnothing$. As a result, the lower λ increases the equilibrium cutoff type. Then, since $m_\varnothing$ reveals $p > \hat{p}$ and m_ω reveals $p < \hat{p}$, the lower λ makes the expressive content of the messages more positive.

A natural question is whether a simple expressive equilibrium is more desirable than the equilibrium without expressive content, i.e., fully directive equilibrium. This is a nontrivial question because of the trade-off between informational and expressive contents; fully directive equilibria are more efficient in terms of communication about ω whereas the expressive message can boost the agent's effort level in simple expressive equilibria. The following result provides useful insight into the trade-off.

Fact 3. *Suppose there exists a simple expressive equilibrium with cutoff $\hat{p}^*$. There exists $\tilde{p} \in (\hat{p}^*,1)$ such that the supervisor with $p > \tilde{p}$ prefers the simple expressive equilibrium*

to a fully directive equilibrium, whereas the supervisor with $p < \tilde{p}$ prefers the fully directive equilibrium to the simple expressive equilibrium.

To see the idea of Fact 3, note that the supervisor's payoff from a directive message in a simple expressive equilibrium is strictly lower than that in a fully directive equilibrium because of the negative expressive content. Thus, when the supervisor sends the directive message, she prefers the fully directive equilibrium to the simple expressive equilibrium. On the other hand, the competent agent's output level conditional on $m_\varnothing$ is higher than the output level in the fully directive equilibrium because of the positive expressive content of $m_\varnothing$. When the agent is incompetent, the output level is lower than that in the fully directive equilibrium since it induces the safe technology. Thus, if p is sufficiently high, the supervisor prefers the simple expressive equilibrium whereas the supervisor with a low p prefers the fully directive equilibrium.

Fact 3 suggests that the supervisor prefers a fully directive equilibrium in the ex ante stage if the probability of $p > \tilde{p}$ is sufficiently small. However, since $\tilde{p}$ depends on the distribution of p, it is not clear whether the supervisor prefers one equilibrium to another in general. The following example shows that the supervisor's favorite equilibrium can be sensitive to the parameter values of the model.

Example 1. Suppose $y(e) = 2e$, $c(e, p) = e^2/2 - pe$ and $\Psi(p)$ is uniform on P. Moreover, $\delta = 0$, $|\Omega| = 10$, and $\pi(\omega)$ is uniform on Ω. If $\lambda = 0.2$, the supervisor prefers a simple expressive equilibrium to a fully directive equilibrium in the ex ante stage. Conversely, she prefers the fully directive equilibrium in the ex ante stage if $\lambda = 0.6$. ◊

As I explained in Fact 2, the expressive content of $m_\varnothing$ is more positive when λ is lower. On the other hand, since a lower λ increases the equilibrium cutoff type, it reduces the ex ante probability of sending $m_\varnothing$. The net effect is not clear in general but, in the case of Example 1, the first effect dominates the second effect.

The agent's favorite equilibrium is also sensitive to the specification and the parameter values of the model. For example, in the setting of Example 1, the competent agent always prefers a simple expressive equilibrium to a fully directive equilibrium whereas the incompetent agent prefers the opposite. Thus, whether the agent prefers one equilibrium to another in the ex ante stage depends on the agent's prior belief. Finally, the supervisor's favorite equilibrium does not always coincide with the agent's favorite equilibrium; for instance, when λ is low in the setting of Example 1, the agent prefers a fully directive equilibrium whereas the supervisor prefers a simple expressive equilibrium in the ex ante stage.

Recall that when there is a state that is very likely to happen, i.e., $(1-\delta)\pi(\omega_{\max}) + \delta > \lambda$, there is no state-independent expressive equilibrium. Thus, an expressive equilibrium has to be state dependent under such situations. Unlike in state-independent expressive equilibria, the supervisor uses a message that can induce a wrong technology to express p in state-dependent expressive equilibria. The next proposition shows that an intuitive expressive equilibrium can exist when there is no simple expressive equilibrium.

PROPOSITION 4. *Suppose* $(1-\delta)\pi(\omega_{\max})+\delta > \lambda$. *If* $\delta < y(e_0)/y(\tilde{e})$, *there exists a consistent expressive equilibrium in which the supervisor uses a communication strategy of the form*

$$\sigma(p,\omega_{\max}) = m_{\max} \quad \textit{for any } p$$
$$\sigma(p,\omega) = \begin{cases} m_{\max} & \textit{if } p \in (\hat{p},1] \\ m(\omega) & \textit{if } p \in [0,\hat{p}), \end{cases}$$

where $m(\omega)$ *is an injection from* Ω *to* $M\backslash\{m_{\max}\}$.

In this equilibrium, the supervisor sends $m_{\max}$ independently of p in $\omega_{\max}$ while she sends $m_{\max}$ in $\omega \neq \omega_{\max}$ only if her trust is higher than a certain level. Moreover, as in simple expressive equilibrium, the supervisor uses a directive message in each $\omega \neq \omega_{\max}$ if her trust is lower than a certain level.

Since $\omega_{\max}$ is very likely to happen, the incompetent agent chooses $a_{\omega_{\max}}$ when the received message is not so informative about ω. As $m_\varnothing$ in a simple expressive equilibrium, $m_{\max}$ expresses her trust by communicating inefficiently about ω. In other words, only the supervisor who strongly believes the agent's competence can send such an uninformative message at $\omega \neq \omega_{\max}$. Unlike $m_\varnothing$ in simple expressive equilibrium, the expressive content of $m_{\max}$ depends on ω. In fact, if the competent agent receives $m_{\max}$ at $\omega_{\max}$, the message has no expressive content, i.e., $P_\sigma(m_{\max},\omega_{\max}) = P$ while it has positive expressive content $P_\sigma(m_{\max},\omega) = (\hat{p},1]$ at any $\omega \neq \omega_{\max}$. When the incompetent agent receives $m_{\max}$, it is not completely uninformative about ω since it is more likely to be sent at $\omega_{\max}$.

The role of the condition $\delta < y(e_0)/y(\tilde{e})$ is similar to that of $\lambda < y(e_0)/y(e_\lambda)$ in simple expressive equilibrium; that is, when this condition is satisfied, inducing a wrong technology is more damaging than making the agent believe $p=0$. When this condition is violated, the conflict of interest in communication about p is too strong to credibly express the supervisor's trust.

Not surprisingly, there can be a state-dependent expressive equilibrium in which the supervisor uses more than two messages in some states. For example, suppose $\Omega = \{1,2,3\}$ and $\omega_{\max} = 1$. Then consider the class of strategies

$$\sigma(p,1) = m_1 \quad \text{for any } p$$
$$\sigma(p,2) = \begin{cases} m_1 & \text{if } p \in (\hat{p}_2,1] \\ m_0 & \text{if } p \in (\hat{p}_1,\hat{p}_2] \\ m_2 & \text{if } p \in [0,\hat{p}_1] \end{cases}$$
$$\sigma(p,3) = \begin{cases} m_1 & \text{if } p \in (\hat{p}_2,1] \\ m_0 & \text{if } p \in (\hat{p}_1,\hat{p}_2] \\ m_3 & \text{if } p \in [0,\hat{p}_1]. \end{cases}$$

If $a(m_1,n) = a_1$ and $a(m_0,n) = a_0$, this can be an equilibrium strategy under some situation. Since m_1 can induce a wrong technology at state 2 and state 3, it handicaps the

supervisor more than m_0 when δ is sufficiently small. Then m_1 conveys a stronger trust than m_0 in equilibrium.

4. Are the pure directives stable?

In fully directive equilibria, the supervisor can communicate about the state efficiently, and all the directives are neutral in expressive content, i.e., pure directives. At first glance, fully directive equilibria seem plausible since there is no conflict of interest in communicating about the state. However, fully directive equilibria might not be so stable. Suppose a supervisor works at an organization where using pure directives are the convention. If she says "I have no suggestion since I trust you," deviating from the convention, she might be able to convey her trust since the focal meaning of the message could induce a suboptimal technology and she would handicap herself. Thus, when the supervisor and the agent happen to share the focal meaning of the off-equilibrium message, it could override the fully directive equilibrium.

To investigate the above observation, I employ the concept of neologism proofness introduced by Farrell (1993). The concept is based on the assumption of a preexisting common language that comes from outside of the game. This assumption makes it possible that the listener can at least "understand" the meaning of a neologism. A putative equilibrium is tested according to whether the supervisor can upset the equilibrium by a *credible neologism*. To define a credible neologism for this game, suppose that when the supervisor's neologism means that her type belongs to $D := Q \times Z \subset P \times \Omega$, the agent updates his belief according to Bayes' rule. That is,

$$\mu_\omega(\omega'|D,n) = \begin{cases} \frac{\pi(\omega')}{\sum_{\omega\in Z}\pi(\omega)} & \text{if } \omega' \in Z \\ 0 & \text{if } \omega' \notin Z \end{cases}$$

$$\mu_p(p'|D,h) = \begin{cases} \frac{\psi(p')}{\int_{p\in Q}\psi(p)\,dp} & \text{if } p' \in Q \\ 0 & \text{if } p' \notin Q \end{cases}$$

for any h. Let $U^D(p,\omega)$ be type (p,ω)'s payoff in which the agent optimally responds to D based on the above updating rule. The term D is a self-signaling set if $(p,\omega) \in D$, $U^D(p,\omega) > U^{\sigma^*(p,\omega)}(p,\omega)$, while $U^D(p,\omega) \leq U^{\sigma^*(p,\omega)}(p,\omega)$ if $(p,\omega) \notin D$. In other words, D is self-signaling when it satisfies a fixed point property: the set of (p,ω) who are strictly better off by claiming she belongs to D exactly coincides with D given an equilibrium. An equilibrium is *neologism-proof* if no self-signaling set exists.

The next result shows that neologism proofness favors simple expressive equilibria over fully directive equilibria.

Proposition 5. *Whenever a simple expressive equilibrium exists, fully directive equilibria are not neologism-proof.*

Whenever a simple expressive equilibrium exists, the supervisor can upset fully directive equilibria by a neologism with $D = (p', 1] \times \Omega$, where p' is higher than the cutoff

type of the simple expressive equilibrium. Intuitively, the neologism expresses her trust while making it credible by being uninformative about ω. Even if no simple expressive equilibrium exists, fully directive equilibria can fail to be neologism-proof. In fact, fully directive equilibria are not neologism-proof whenever the safe choice is the incompetent agent's optimal choice under $\pi(\omega)$. Note that this condition is always satisfied when a simple expressive equilibrium exists.

To rank simple expressive and fully directive equilibria based on neologism proofness, I need to show that simple expressive equilibria can be neologism-proof under a reasonable condition. It can be shown that if a simple expressive equilibrium is not neologism-proof, any self-signaling set needs to take the form $D = [0, q) \times \{\omega\}$, where $q > \hat{p}^*$.[11] To obtain further results, I specify the disutility function as $c(e, p; \beta) = C(e) - \beta b(e, p)$, where $\beta \in (0, 1]$. One interpretation of this specification is that the agent's net disutility of effort $c(e, p)$ is determined by his "pain," i.e., $C(e)$, and his "enjoyment" that can be enhanced by the supervisor's trust, i.e., $\beta b(e, p)$. Then assume that $C : E \to \Re_+$ is twice differentiable and strictly convex whereas $b : E \times P \to \Re_+$ is twice differentiable in each argument and $b_1(e, p) > 0$. Moreover, assume $b_{11}(e, p) \leq 0$, $b_{12}(e, p) > 0$, and $C'(e) - b_1(e, p) > 0$ for any (e, p) so that the assumptions about $c(e, p)$ in Section 2 are guaranteed under any $\beta \in (0, 1]$. The following result states that a simple expressive equilibrium is neologism-proof if the marginal disutility of effort is sufficiently insensitive to p.

PROPOSITION 6. *If β is sufficiently small, any simple expressive equilibrium is neologism-proof.*

The idea of Proposition 6 is as follows. As I mentioned earlier, whenever a credible neologism exists, the self-signaling set takes the form of $[0, q) \times \{\omega\}$, where $q > \hat{p}^*$. Note that the supervisor's payoff from $D = [0, q) \times \{\omega\}$ with $q = \hat{p}^*$ is the same as the cutoff type's payoff in the simple expressive equilibrium with $\hat{p}^*$. When β is lower, the agent's effort becomes less sensitive to q. Then, since the supervisor's payoff from $m_\varnothing$ is strictly increasing in p, the neologism with any $q > \hat{p}^*$ cannot be more attractive than sending $m_\varnothing$ in the simple expressive equilibrium. How small is "sufficiently small" when a simple expressive equilibrium is neologism-proof? The next result shows that the condition in Proposition 6 can be quite mild.

FACT 4. *Suppose $\Psi(p)$ is a uniform distribution on $[0, 1]$, $y(e) = \alpha e$, and $c(e, p) = e^2/2 - \beta p e$, where $\alpha > \beta > 0$. There exists a simple expressive equilibrium if $(1 - \delta)\pi(\omega_{\max}) + \delta \leq \lambda$ and $\beta < 2(1 - \lambda^2)\alpha/\lambda$. Moreover, whenever a simple expressive equilibrium exists, it is neologism-proof.*

Fact 4 says that whenever β supports a simple expressive equilibrium in the linear-quadratic-uniform setting, it is always small enough to make the equilibrium neologism-proof.

[11] For more details, see the proof of Proposition 5 in the Appendix.

When there are multiple simple expressive equilibrium outcomes, only those with the highest cutoff can be neologism-proof.[12] The idea of this result is as follows. If there is another simple expressive equilibrium with a higher cutoff type, the supervisor's expected payoff from neologism $[0, q) \times \{\omega\}$ is sensitive to q. As a result, a higher q makes the payoff from the neologism higher than type q's equilibrium payoff. Since type q's equilibrium payoff becomes higher than that from the neologism as q goes to 1, there exists q that makes $[0, q) \times \{\omega\}$ a self-signaling set. Note that this result does not contradict Proposition 6; there exists a unique simple expressive equilibrium outcome if β is sufficiently low.

Remark 3. While neologism proofness turned out to be effective to rank the two important equilibria, the fixed point property of self-signaling sets makes obtaining more general results difficult. To see this, consider any consistent expressive equilibrium. While neologism $[0, q) \times \{\omega\}$ is a typical candidate for a credible neologism, whether it can be a self-signaling set or not depends on the sensitivity of the disutility function to p; that is, it is a quantitative question that depends on the specification and the parameter values of the model. A disutility function that is less sensitive to p can make such a neologism noncredible. However, the class of expressive equilibria could fail to exist when a disutility function becomes less sensitive to p.

5. Discussion

5.1 *Two-way communication*

This paper focuses on one-way communication. The setting captures a situation where there is a formal or informal rule that does not allow the agent to speak first. Another interpretation is that the agent has no incentive to reveal his type because of some factor that is not in the model. For example, suppose the supervisor could replace the agent before the task if she knows the agent is incompetent. Then, since the incompetent agent has no incentive to reveal his type, we can analyze the communication problem as one-way communication without loss of generality.

However, one might ask how the nature of communication could be changed if the agent can speak first, i.e., two-way communication. To investigate the question, the assumption of "belief-dependent disutility" needs to be clarified for two-way communication since the supervisor could update her belief based on the agent's message. Recall that the assumption is based on the idea that distrust reduces motivation. However, when the agent voluntarily reveals his competence in equilibrium, there is no room for distrust. Thus, if the agent claims that he is incompetent, it should not be offensive that the supervisor perceives the agent as incompetent. Hence, it seems natural to stay with the assumption that the agent's disutility depends on the supervisor's prior p, i.e., her personal view about the agent's competence, rather than her posterior belief that is based on the agent's message.

First, consider a communication protocol that forces the agent to speak first. Specifically, in the first stage, the agent sends a costless message about his competence to the

[12] See the Appendix for the proof of this result.

supervisor. In the second stage, the supervisor sends $m \in M$ given the agent's message and (p, ω). In the third stage, the agent makes decisions given m, and the game ends. Since the agent's cheap talk in the first stage can be ignored, this setting expands the set of equilibria in the basic setting. However, this setting has a notable feature: there is a neologism-proof equilibrium that is payoff-equivalent to a fully directive equilibrium. In this equilibrium, the agent's message reveals his competence, and the supervisor sends a directive message only if the agent is incompetent. Since the supervisor learns the agent's competence in the first stage, the supervisor has no room to use a neologism to upset the equilibrium.

Whether the supervisor prefers the above two-way communication protocol to the one-way communication in the basic setting is not clear. Note that Example 1 shows that whether the supervisor prefers a fully directive equilibrium to a simple expressive equilibrium in the ex ante stage depends on the parameter values. Moreover, even in the case in which the supervisor prefers a fully directive equilibrium in the ex ante stage, Fact 3 shows that the supervisor with $p > \tilde{p}$ always prefers a simple expressive equilibrium in the interim stage. In other words, the supervisor with a high p might try to signal her trust by refusing the two-way communication and speaking one-directionally. This observation calls for the following "turn-taking" model.

Suppose that the supervisor decides the communication protocol given (p, ω), and the agent can speak only if the supervisor allows it. The other aspects of the model is the same as in the basic setting.

- Stage 1. The supervisor chooses either sending a message one-directionally or letting the agent speak first. Formally, her strategy is $\sigma_1 : P \times \Omega \to M \cup \{ask\}$.
- Stage 2.
  - If $\sigma_1(p, \omega) = m$, then the agent chooses actions (a, e) and the game ends; that is, it is the same as in the basic setting.
  - If $\sigma_1(p, \omega) = ask$, the agent reports his competence $r \in \{C, I\}$; that is, his strategy is $r : T \to \{C, I\}$. Then the game moves to Stage 3.
- Stage 3. The supervisor chooses $m \in M$; that is, her strategy is $\sigma_2 : P \times \Omega \times \{C, I\} \to M$.
- Stage 4. Given message m, the agent chooses actions (a, e) and the game ends.

Interestingly, this two-way communication game has equilibria that are analogous to those in the original one-way communication game.

FACT 5. *Whenever a simple expressive equilibrium exists in the basic setting, there exists an equilibrium in which*

$$\sigma_1(p, \omega) = \begin{cases} m_\varnothing & \text{if } p > \hat{p}^* \\ ask & \text{if } p < \hat{p}^* \end{cases}$$

$$r(t) = \begin{cases} C & \text{if } t = C \\ I & \text{if } t = I \end{cases}$$

$$\sigma_2(p, \omega, r) = \begin{cases} m(\omega) & \text{if } r = I \\ m_\varnothing & \text{if } r = C, \end{cases}$$

where $\hat{p}^*$ is the cutoff type of the simple expressive equilibrium. Moreover, this equilibrium is payoff-equivalent to the simple expressive equilibrium in the basic game.

Note that when $p > \hat{p}^*$, the agent's belief conditional on $m_\varnothing$ is the same as that in the simple expressive equilibrium with cutoff $\hat{p}^*$. Thus, the supervisor's payoff from $m_\varnothing$ is the same as that in the simple expressive equilibrium. On the other hand, if $p < \hat{p}^*$, she asks the agent's type and gives advice on request. However, since her turn-taking decision reveals her distrust, i.e., $p < \hat{p}^*$, and the agent knows ω in the final stage, the agent's decision is the same as the case in which he receives a directive message in the simple expressive equilibrium. Thus, this equilibrium is payoff-equivalent to the simple expressive equilibrium. Intuitively, the fact that the supervisor questions the agent's competence reveals her distrust while the supervisor can express her trust by not asking any question.

There also exists an equilibrium that is payoff-equivalent to a fully directive equilibrium. In this equilibrium, the supervisor asks about the agent's competence independently of her type; the agent truthfully reports his competence; the supervisor sends a directive message only if the agent is incompetent. However, this equilibrium suffers from the same problem as fully directive equilibrium in the basic setting: this is not neologism-proof whenever a simple expressive equilibrium exists. Thus, if we consider two-way communication in which the supervisor is not a passive communicator, the basic nature of the model goes back to the original setting; that is, the basic setting provides useful insight into two-way communication.

5.2 *A related signaling game and the D1 equilibrium*

Section 4 demonstrated that neologism proofness favors simple expressive equilibria over fully directive equilibria. One might think that the instability of pure directives could be the product of neologism-proof equilibrium, which is known to be a stringent criterion. I claim that as long as the supervisor can handicap herself with an off-equilibrium message, we can obtain the analogous result even under a different type of off-equilibrium belief restriction.

To see the claim, suppose $M = A$. Since there is no conflict of interest in communication about ω, suppose that the incompetent's choice function is restricted so that $a(m, n) = m$. This restriction transforms the current model into a costly signaling model with single type p. As a result, some equilibrium dominance-based refinement becomes effective. In the Appendix, it is shown that the fully directive equilibrium fails the D1 criterion whenever the simple expressive equilibrium exists. Since the simple expressive

equilibrium always passes the test, the result is analogous to that with neologism-proof equilibrium.[13]

5.3 *Misunderstanding*

In equilibrium, there is no misunderstanding about the expressive content of directives. However, in reality, we often suffer from misunderstanding in expressive content. For example, when the supervisor says "Choose technology 1," the agent could interpret the message with negative expressive content even if the supervisor does not intend it. In fact, linguistic philosophers treat a speaker's intended effect of her utterance, i.e., illocutionary force, and the actual effect, i.e., the perlocutionary effect, separately.

One way to deal with the gap between the speaker's intention and the listener's interpretation might be to incorporate some naive/sincere types who use words literally as Chen (2011). Another approach might be to incorporate "language competence" as do Blume and Board (2013). Their approach is to introduce "language type" that determines the set of messages one can send and understand. However, since the relevant language competence here is not in the ability to send or identify messages but "to play equilibrium," it could be challenging to extend their approach.

6. Concluding remarks

Since the expressive content of words could affect economic decisions, it is important to understand how the expressive content could emerge in an economic context. This paper provided a game theoretical framework to investigate how a seemingly uninformative message could have positive expressive content while directives could have negative expressive content. It is also shown that pure directives are susceptible to an intuitive neologism whenever an expressive equilibrium exists.

Even though the model of this paper is built on a psychological setting, i.e., the belief-dependent payoff, the idea of the trade-off between expressives and directives can be applicable to standard settings. For example, consider a situation in which a worker needs to choose his effort level for a task but it also has an aspect of relationship-specific investment. Since the agent's investment decision could depend on how the supervisor thinks about the agent's ability, the manager might need to balance the informational and expressive contents of her directives.

Appendix

A.1 *Proof of Proposition 1*

Let $U_\omega(p)$ be the supervisor's payoff at ω given p in an equilibrium, that is,

$$U_\omega(p) = py(e(\sigma(p,\omega),\omega)) + (1-p)f(a(\sigma(p,\omega),n), e(\sigma(p,\omega),n),\omega).$$

First, I establish the following lemma.

[13] I appreciate one of the referees for motivating this observation.

LEMMA 3. *The function $U_\omega(p)$ is a piecewise linear convex function.*

PROOF. First, I claim that $U_\omega(p)$ is a continuous function. Clearly, $U_\omega(p)$ is continuous at p' if $p' \in \text{int}(P_\sigma(\sigma(p',\omega),\omega))$. So suppose $p' \in \partial P_\sigma(m',\omega)$ and that $U_\omega(p)$ is discontinuous at p', that is, $U_\omega(p') \neq \lim_{p\to p'} U_\omega(p)$. Let $U^m(p,\omega)$ be type p's payoff from equilibrium message m at ω in the equilibrium. Note that given any $\tilde{p} \in P$, $\lim_{p\to\tilde{p}} U^{\sigma(\tilde{p},\omega)}(p,\omega) = U_\omega(\tilde{p})$. Thus, if $U_\omega(p') > \lim_{p\to p'} U_\omega(p)$, $U^{\sigma(p',\omega)}(p'',\omega) > U_\omega(p'')$ for some $p'' \notin P_\sigma(\sigma(p',\omega),\omega)$ such that $p' \in \partial P_\sigma(\sigma(p'',\omega),\omega)$ and that is close to p'. But then p'' has an incentive to send m'. On the other hand, if $U_\omega(p') < \lim_{p\to p'} U_\omega(p)$, p' has an incentive to imitate some $p \notin P_\sigma(\sigma(p',\omega),\omega)$ such that $p' \in \partial P_\sigma(\sigma(p,\omega),\omega)$ and that is close to p'. Thus, $U_\omega(p)$ is continuous in p.

Second, $U_\omega(p)$ is piecewise linear since, for any $p \in P_\sigma(m,\omega)$, $U'_\omega(p;m) = y(e(m,\omega)) - f(e(m,n),a(m,n),\omega)$, which is constant in $p \in P_\sigma(m,\omega)$, and $P_\sigma(m,\omega)$ is a partition of P.

Finally, since $U_\omega(p)$ is piecewise linear and continuous, it suffices to show that $U'_\omega(p)$ is weakly increasing to complete the proof. Suppose $U'_\omega(p)$ is not weakly increasing. Then, since the function is continuous and piecewise linear, there exist (p_0,p_1,p_2) such that (i) $p_1 < p_0 < p_2$, (ii) $p_0 \in \partial P_\sigma(\sigma(p_1,\omega),\omega) \cap \partial P_\sigma(\sigma(p_2,\omega),\omega)$, and (iii) $k_\sigma(\sigma(p_2,\omega),\omega) < k_\sigma(\sigma(p_1,\omega),\omega)$. Observe that

$$U^{m''}(p_2,\omega) = U^{m''}(p_0,\omega) + \int_{p_0}^{p_2} k_\sigma(m'',\omega)\,dp$$

$$U^{m'}(p_2,\omega) = U^{m'}(p_0,\omega) + \int_{p_0}^{p_2} k_\sigma(m',\omega)\,dp,$$

where $m' = \sigma(p_1,\omega)$ and $m'' = \sigma(p_2,\omega)$. Since $U^{m'}(p_0,\omega) = U^{m''}(p_0,\omega)$, $U^{m'}(p_2,\omega) > U^{m''}(p_2,\omega)$ if $k_\sigma(m'',\omega) < k_\sigma(m',\omega)$. But then p'' strictly prefers m' to m'', a contradiction. □

The proof of Proposition 1 consists of three steps.

Step 1. If $e(m',\omega_1) < e(m',n)$ in an influential expressive equilibrium, there exists $\omega_2 \neq \omega_1$ such that $e(m',\omega_2) > e(m',n)$. Let $\Omega_\sigma(m) = \{\omega : \sigma(p,\omega) = m \text{ for some } p\}$. Note that

$$\mu_p(p|m,n) = \sum_{\omega\in\Omega_\sigma(m)} \mu_p(p|m,\omega)\pi_\sigma(\omega|m),$$

where $\pi_\sigma(\omega|m) = \pi(\omega)/\sum_{\tilde{\omega}\in\Omega_\sigma(m)} \pi(\tilde{\omega})$. Since $e(m',\omega_1) < e(m',n)$ while $y(e)$ is concave and $y'(e) \geq f_1(e,a,\omega)$ given any a and ω, the incompetent agent's expected marginal disutility from $e(m',n)$ conditional on m' needs to be lower than the competent's, that is,

$$\int_p c_1(e(m',n),p) \sum_{\omega\in\Omega_\sigma(m)} \mu_p(p|m,\omega)\pi_\sigma(\omega|m)\,dp \leq \int_p c_1(e(m',\omega_1),p)\mu_p(p|m,\omega_1)\,dp.$$

Since $c_1(e,p) > 0$ given any p and $e(m',\omega_1) < e(m',n)$, the competent agent's expected marginal disutility from $e(m',\omega_1)$ is strictly lower than that from $e(m',n)$, that

is,

$$\int_p c_1(e(m', \omega_1), p)\mu_p(p|m, \omega_1)\, dp < \int_p c_1(e(m', n), p)\mu_p(p|m, \omega_1)\, dp.$$

From the above two inequalities,

$$\int_p c_1(e(m', n), p) \sum_{\omega \in \Omega_\sigma(m)} \mu_p(p|m, \omega)\pi_\sigma(\omega|m)\, dp < \int_p c_1(e(m', n), p)\mu_p(p|m, \omega_1)\, dp.$$

Since $\omega_1 \in \Omega_\sigma(m)$, there must exist $\omega_2 \in \Omega_\sigma(m)$ such that

$$\int_p c_1(e(m', n), p) \sum_{\omega \in \Omega_\sigma(m)} \mu_p(p|m, \omega)\pi_\sigma(\omega|m)\, dp > \int_p c_1(e(m', n), p)\mu_p(p|m, \omega_2)\, dp.$$

Then, since $f_1(e(m', n), a(m', n), \omega_2) \leq y'(e(m', n))$,

$$y'(e(m', n)) > \int_p c_1(e(m', n), p)\mu_p(p|m', \omega_2)\, dp.$$

Since $y(e)$ is concave and the expected marginal disutility conditional on m' is strictly increasing, $e(m', \omega_2)$ that satisfies the first-order condition has to be strictly higher than $e(m', n)$.

Step 2. In any influential expressive equilibrium, $U'_\omega(p) > 0$ for some (p, ω). Note that $U'_\omega(p; m) = y(e(m, \omega)) - f(e(m, n), a(m, n), \omega)$. If $e(\sigma(p, \omega), \omega) > e(\sigma(p, \omega), n)$ for some (p, ω), then $U'_\omega(p) > 0$ since $y(e) \geq f(e, a, \omega)$ for all e given any a and ω. On the other hand, from Step 1, if $e(\sigma(p, \omega), \omega) < e(\sigma(p, \omega), n)$ for some (p, ω), then $e(\sigma(p, \omega), \omega') > e(\sigma(p, \omega), n)$ for some $\omega' \neq \omega$, and thus $U'_{\omega'}(p) > 0$. Now suppose $e(\sigma(p, \omega), \omega) = e(\sigma(p, \omega), n)$ and $U'_\omega(p) = 0$ for all (p, ω). Then $\sigma(p, \omega)$ is directive for any (p, ω). If this is an influential equilibrium, $e(m'', \omega) > e(m', \omega)$ for some m'' and m'. However, if both messages are directive, no p prefers m' to m'', a contradiction.

Step 3. There exists (p', ω') such that $\sigma(p, \omega')$ is not directive for all $p \in (p', 1]$. From Lemma 3 and Step 2, there exists ω' such that $U'_{\omega'}(1) > 0$. Since $U'_{\omega'}(1) \neq 0$, $\sigma(1, \omega')$ is not directive. Then choose p' so that $(p', 1] \subset P_\sigma(\sigma(1, \omega'), \omega')$.

A.2 *Proof of Lemma 1*

Let $m' = \sigma(p', \omega')$ and $m'' = \sigma(p'', \omega')$. Moreover, let $Y(m, \omega)$ ($Y(m, n)$) be the equilibrium output of the competent (incompetent) agent, respectively. From the equilibrium conditions,

$$p'[Y(m'', \omega') - Y(m', \omega')] + (1 - p')[Y(m'', n) - Y(m', n)] \leq 0$$
$$p''[Y(m'', \omega') - Y(m', \omega')] + (1 - p'')[Y(m'', n) - Y(m', n)] \geq 0.$$

Then, since $p'' > p'$, $Y(m'', \omega') - Y(m', \omega') \geq 0$ and $Y(m'', n) - Y(m', n) \leq 0$. From $Y(m'', \omega') - Y(m', \omega') \geq 0$, $e(m'', \omega') \geq e(m', \omega')$.

Now, suppose m'' is directive but m' is not. Since $Y(m'', n) - Y(m', n) \leq 0$, and $y(e) \geq f(e, a, \omega)$ for all e given any a and ω, then $e(m'', n) \leq e(m', n)$. Since m'' is directive,

$e(m'', n) = e^c(m'', n)$. On the other hand, since m' is not directive, $e(m', n) < e^c(m', n)$. But then $e^c(m'', n) < e^c(m', n)$, that is, this is not consistent expressive equilibrium.

A.3 *Proof of Lemma 2*

Let $M_\sigma(\omega) = \{m : \sigma(p, \omega) = m \text{ for some } p\}$. In any state-independent expressive equilibrium, $\Omega_\sigma(m) = \Omega$ or $\{\omega\}$ for any $m \in M_\sigma(\omega)$. Thus, if some message makes the incompetent agent choose a technology that is optimal at another state, there are two cases.

First, suppose that $\Omega_\sigma(m) = \Omega$ for all $m \in M_\sigma(\omega')$. Then if $a(m, n) = a_{\omega''}$ for some $m \in M_\sigma(\omega')$, then $a(m, n) = a_{\omega''}$ for all $m \in M_\sigma(\omega')$. Since this is an influential expressive equilibrium, $e(m', \omega') > e(m'', \omega')$ for some $m', m'' \in M_\sigma(\omega')$. Then, since $\mu_\omega(\omega|m'', n) = \mu_\omega(\omega|m', n)$ while $\mu_p(p|m, \omega') = \mu_p(p|m, n)$ for $m = m', m''$, $f(e(m', n), a(m', n), \omega') > f(e(m'', n), a(m'', n), \omega')$. But then no p has an incentive to sends m'' at ω', a contradiction.

Second, suppose $\Omega_\sigma(m) = \{\omega'\}$ for some $m \in M_\sigma(\omega')$. Then, from Lemma 1, there exist $m', m'' \in M_\sigma(\omega')$ such that $a(m', n) = a_{\omega''}$, $a(m'', n) = a_{\omega'}$, and $\inf P_\sigma(m', \omega') = \sup P_\sigma(m'', \omega')$. Since $\Omega_\sigma(m') = \Omega$ and $a(m', n) = a_{\omega''}$, $U_{\omega'}(p) < U_{\omega''}(p)$ for all $p \in P_\sigma(m', \omega')$ form the state-independence property. On the other hand, from Lemma 1 and the state-independence property, $\Omega_\sigma(\sigma(p, \omega)) = \{\omega\}$ for any $p < p'$.

I claim that $e(\sigma(p, \omega'), \omega') = e(\sigma(p, \omega''), \omega'')$ for any $p < p'$. Since $p < p'$ sends a directive message, from Lemma 3, $e(\sigma(p, \omega), \omega)$ is constant in p for $p < p'$. Let $e_\omega = e(\sigma(p, \omega), \omega)$ for $p < \hat{p}$ and $M_\sigma(\omega; p < \hat{p}) = \{m : \sigma(p, \omega) = m \text{ for some } p < \hat{p}\}$. Note that $y'(e_\omega) = E_p[c_1(e_\omega, p)|P_\sigma(m, \omega), \omega]$ for any $m \in M_\sigma(\omega; p < \hat{p})$. Thus, we can write

$$E_p[c_1(e_\omega, p)|P_\sigma(m', \omega), \omega] = \sum_{m \in M_\sigma(\omega; p < \hat{p})} \Pr(p \in P_\sigma(m, \omega)|p < \hat{p}) E_p[c_1(e_\omega, p)|P_\sigma(m, \omega), \omega]$$

for any $m' \in M_\sigma(\omega; p < \hat{p})$. Thus,

$$\begin{aligned} &E_p[c_1(e_\omega, p)|P_\sigma(m', \omega), \omega] \\ &\quad = \sum_{m \in M_\sigma(\omega; p < \hat{p})} \frac{\Pr(p \in P_\sigma(m, \omega))}{\Psi(\hat{p})} \int_{p \in P_\sigma(m, \omega)} c_1(e_\omega, p) \frac{\psi(p)}{\Pr(p \in P_\sigma(m, \omega))} dp \\ &\quad = \int_{p < \hat{p}} c_1(e_\omega, p) \frac{\psi(p)}{\Psi(\hat{p})} dp. \end{aligned}$$

Thus, $e_\omega = e'$, where $y'(e') = E[c_1(e', p)|p < \hat{p}]$, that is, e_ω is independent of ω. Then $U_{\omega'}(p) = U_{\omega''}(p)$ for all $p < p'$. However, since $U_{\omega'}(p') < U_{\omega''}(p')$, $U_{\omega''}(p) < U_{\omega''}(p')$ for all $p < p'$ whenever $U_{\omega'}(p) = U_{\omega'}(p')$ for all $p < p'$, violating Lemma 3.

A.4 *Proof of Proposition 2*

As I showed in the proof of Lemma 2, in any state-independent expressive equilibrium, $\Omega_\sigma(\sigma(p, \omega)) = \{\omega\}$ for some (p, ω). From Proposition 1, there is no influential expressive equilibrium in which all messages are directive. Thus, $\Omega_\sigma(\sigma(p, \omega)) = \Omega$ for some

(p, ω). Moreover, since any state-independent expressive equilibrium is a consistent expressive equilibrium, Lemma 1 implies that if $\sigma(p', \omega)$ is directive, $\sigma(p, \omega)$ is also directive for any $p < p'$. Then, from Lemma 2, given a state-independent expressive equilibrium, there exists $\hat{p}$ such that $\sigma(p, \omega)$ is directive if $p < \hat{p}$ whereas $a(\sigma(p, \omega), \omega) = a_0$ for any $p > \hat{p}$. Moreover, by the state-independence property, such $\hat{p}$ is independent of ω.

Now I claim that any state-independent expressive equilibrium that is characterized by $\hat{p}$ is payoff-equivalent to a simple expressive equilibrium with $\hat{p}$. Let σ^{se} be a simple expressive strategy with $\hat{p}$, and let $U^{se}_{\omega}(p)$ be type p's payoff from σ^{se} at ω when the agent responds optimally.

To establish the claim, first I show $U_{\omega}(p) = U^{se}_{\omega}(p)$ for $p > \hat{p}$. From the state-independence property, if $e(\sigma(p', \omega), \omega) \geq (<) e(\sigma(p'', \omega), \omega)$, then $e(\sigma(p', \omega), n) \geq (<) e(\sigma(p'', \omega), n)$. Thus, if $e(\sigma(p', \omega), \omega) > (<) e(\sigma(p'', \omega), \omega)$, then p'' (p') imitates p' (p''). Hence, $e(\sigma(p', \omega), \omega) = e(\sigma(p'', \omega), \omega)$ for any $p', p'' > \hat{p}$ in the equilibrium. Then let $e' = e(\sigma(p, \omega), \omega)$ for $p > \hat{p}$, and $M_{\sigma}(\omega; p > \hat{p}) = \{m : \sigma(p, \omega) = m \text{ for some } p > \hat{p}\}$. Since $y'(e') = E_p[c_1(e', p) | P_{\sigma}(m, \omega), \omega]$ for any $m \in M_{\sigma}(\omega; p > \hat{p})$, we can write

$$E_p[c_1(e', p) | P_{\sigma}(m', \omega), \omega] = \sum_{m \in M_{\sigma}(\omega; p > \hat{p})} \Pr(p \in P_{\sigma}(m, \omega) | p > \hat{p}) E_p[c_1(e', p) | P_{\sigma}(m, \omega), \omega]$$

for any $m' \in M_{\sigma}(\omega; p > \hat{p})$. Thus,

$$\begin{aligned} &E_p[c_1(e', p) | P_{\sigma}(m', \omega), \omega] \\ &= \sum_{m \in M_{\sigma}(\omega; p > \hat{p})} \frac{\Pr(p \in P_{\sigma}(m, \omega))}{1 - \Psi(\hat{p})} \int_{p \in P_{\sigma}(m, \omega)} c_1(e', p) \frac{\psi(p)}{\Pr(p \in P_{\sigma}(m, \omega))} dp \\ &= \sum_{m \in M_{\sigma}(\omega; p > \hat{p})} \int_{p \in P_{\sigma}(m, \omega)} c_1(e', p) \frac{\psi(p)}{1 - \Psi(\hat{p})} dp \\ &= \int_{p > \hat{p}} c_1(e', p) \frac{\psi(p)}{1 - \Psi(\hat{p})} dp. \end{aligned}$$

Then, since e' satisfies the first-order condition $y'(e') = E[c_1(e', p) | p > \hat{p}]$, $e' = e(\sigma^{se}(p, \omega), \omega)$ for all $p > \hat{p}$. Then, from the state-independence property, $U_{\omega}(p) = U^{se}_{\omega}(p)$ for $p > \hat{p}$.

Now I show $U_{\omega}(p) = U^{se}_{\omega}(p)$ for $p < \hat{p}$. As I mentioned earlier, $e(\sigma(p', \omega), \omega) = e(\sigma(p'', \omega), \omega)$ for any $p', p'' < \hat{p}$. Let $e'' = e(\sigma(p, \omega), \omega)$ for $p < \hat{p}$ and $M_{\sigma}(\omega; p < \hat{p}) = \{m : \sigma(p, \omega) = m \text{ for some } p < \hat{p}\}$. Note that $y'(e'') = E_p[c_1(e'', p) | P_{\sigma}(m, \omega)]$ for any $m \in M_{\sigma}(\omega; p < \hat{p})$. Thus, by an argument analogous to the case of $p > \hat{p}$, $E_p[c_1(e'', p) | P_{\sigma}(m, \omega)] = E[c_1(e'', p) | p < \hat{p}]$ for all $m \in M_{\sigma}(\omega; p < \hat{p})$. Then, since e'' satisfies $y'(e'') = E[c_1(e'', p) | p < \hat{p}]$, $e'' = e(\sigma^{se}(p, \omega), \omega)$ for all $p < \hat{p}$. Thus, $U_{\omega}(p) = U^{se}_{\omega}(p)$ for $p < \hat{p}$.

Since the case of $p = \hat{p}$ can be included in either $p > \hat{p}$ or $p < \hat{p}$, $U_{\omega}(p) = U^{se}_{\omega}(p)$ for all p.

A.5 *Proof of Proposition 3*

Suppose that the supervisor uses a simple expressive strategy with cutoff $\hat{p}$. Let $e(m, h|\hat{p})$ be the optimal effort level given (m, h) when the supervisor uses the simple expressive strategy with $\hat{p}$. The type (p, ω) supervisor's expected payoff from $m_\varnothing$ is then

$$U^{m_\varnothing}(p, \omega|\hat{p}) = py(e(m_\varnothing, \omega|\hat{p})) + (1-p)f(e(m_\varnothing, n|\hat{p}), a(m_\varnothing, n), \omega).$$

Note that the incompetent agent's expected payoff given a_0 and $m_\varnothing$ is

$$\max_{e\in E}\left\{\lambda y(e) - \int_p c(e, p)\mu_p(p|m_\varnothing, n)\,dp\right\}.$$

Moreover, his payoff given $a_{\omega'}$ and $m_\varnothing$ is

$$\max_{e\in E}\left\{\pi(\omega')y(e) + (1-\pi(\omega'))\delta y(e) - \int_p c(e, p)\mu_p(p|m_\varnothing, n)\,dp\right\}.$$

Thus, if $\lambda \geq (1-\delta)\pi(\omega) + \delta$ for any ω, then a_0 is optimal for the incompetent agent given $m_\varnothing$. The type (p, ω) supervisor's expected payoff from $m_\varnothing$ is then

$$U^{m_\varnothing}(p, \omega|\hat{p}) = py(e(m_\varnothing, \omega|\hat{p})) + (1-p)\lambda y(e(m_\varnothing, n|\hat{p})).$$

Since $\mu_\omega(\omega|m_\omega, n) = 1$, we have $e(m_\omega, \omega|\hat{p}) = e(m_\omega, n|\hat{p})$. Thus, the supervisor's expected payoff from m_ω is

$$U^{m_\omega}(p, \omega|\hat{p}) = y(e(m_\omega, \omega|\hat{p})).$$

Given the simple expressive strategy with $\hat{p}$, the supervisor sends $m_\varnothing$ only if $p > \hat{p}$. Thus, the agent's consistent belief about p conditional on $m_\varnothing$ is

$$\mu_p(p|m_\varnothing, h; \hat{p}) = \begin{cases} \frac{\psi(p)}{1-\Psi(\hat{p})} & \text{if } p \in (\hat{p}, 1] \\ 0 & \text{if } p \in [0, \hat{p}] \end{cases}$$

for any h. Henceforth, I omit h from $\mu_p(p|m_\varnothing, h; \hat{p})$.

Note that $\mu_p(p|m_\varnothing; \hat{p})$ is not well defined at $\hat{p} = 1$. Then consider the limit of $\mu_p(p|m_\varnothing; \hat{p})$ when $\hat{p}$ approaches 1. Note that $\operatorname{supp}(\mu_p(\cdot|m_\varnothing; \hat{p})) = (\hat{p}, 1]$ and $\mu_p(1|m_\varnothing; \hat{p}) > 0$ for any $\hat{p} < 1$. Thus, as $\hat{p}$ approaches 1, the only possible type who sends $m_\varnothing$ becomes $p = 1$. Thus,

$$\lim_{\hat{p}\to 1}\int_{\hat{p}}^{1} c(e, p)\mu_p(p|m_\varnothing; \hat{p})\,dp = c(e, 1).$$

Then, by the maximum theorem, $\lim_{\hat{p}\to 1} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) = y(e_1)$, where

$$e_1 := \arg\max_{e\in E}\{y(e) - c(e, 1)\}.$$

Given the simple expressive strategy with $\hat{p}$, the supervisor sends m_ω only if $p \le \hat{p}$. Thus, the agent's consistent belief about p conditional on m_ω is

$$\mu_p(p|m_\omega, h; \hat{p}) = \begin{cases} \frac{\psi(p)}{\Psi(\hat{p})} & \text{if } p \in [0, \hat{p}] \\ 0 & \text{if } p \in (\hat{p}, 1] \end{cases}$$

for any h. Henceforth, I omit h from $\mu_p(p|m_\omega, h; \hat{p})$.

Obviously, $\lim_{\hat{p}\to 1} \mu_p(p|m_\omega; \hat{p}) = \mu_p(p|m_\omega; 1) = \psi(p)$. Then, since $a(m_\omega, h) = a_\omega$ for $h = \omega, n$, $\lim_{\hat{p}\to 1} U^{m_\omega}(\hat{p}, \omega|\hat{p}) = U^{m_\omega}(1, \omega|1) = y(\tilde{e})$, where

$$\tilde{e} = \arg\max_{e\in E}\left\{ y(e) - \int_p c(e, p)\psi(p)\, dp \right\}.$$

Then, since $e_1 > \tilde{e}$,

$$\lim_{\hat{p}\to 1} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) > \lim_{\hat{p}\to 1} U^{m_\omega}(\hat{p}, \omega|\hat{p}).$$

Turning to the case where $\hat{p}$ approaches 0, note that $\lim_{\hat{p}\to 0} \mu(p|m_\varnothing; \hat{p}) = \mu(p|m_\varnothing; 0) = \psi(p)$. Then, since $m_\varnothing$ induces the incompetent agent to choose a_0, $\lim_{\hat{p}\to 0} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) = U^{m_\varnothing}(0, \omega|0) = \lambda y(e_\lambda)$, where

$$e_\lambda = \arg\max_{e}\left\{ \lambda y(e) - \int_p c(e, p)\psi(p)\, dp \right\}.$$

Consider the case of m_ω. Since $\mu_p(p|m_\omega; \hat{p})$ is not well defined at $\hat{p} = 0$, we need to consider the limit of $\mu_p(p|m_\omega; \hat{p})$ when $\hat{p}$ approaches 0. Note that $\mathrm{supp}(\mu_p(\cdot|m_\omega; \hat{p})) = [0, \hat{p}]$ and $\mu_p(0|m_\omega; \hat{p}) > 0$ for any $\hat{p} > 0$. Thus, as $\hat{p}$ approaches 0, the only possible type who sends m_ω becomes $p = 0$. Hence,

$$\lim_{\hat{p}\to 0} \int_0^{\hat{p}} c(e, p)\mu_p(p|m_\omega; \hat{p})\, dp = c(e, 0).$$

Then, by the maximum theorem, $\lim_{\hat{p}\to 0} U^{m_\omega}(\hat{p}, \omega|\hat{p}) = y(e_0)$, where

$$e_0 = \arg\max_{e}\{y(e) - c(e, 0)\}.$$

Therefore, if $\lambda y(e_\lambda) < y(e_0)$,

$$\lim_{\hat{p}\to 0} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) < \lim_{\hat{p}\to 0} U^{m_\omega}(\hat{p}, \omega|\hat{p}).$$

Since $U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) - U^{m_\omega}(\hat{p}, \omega|\hat{p})$ is continuous in $\hat{p} \in (0, 1)$, there exists at least one $\hat{p}^* \in (0, 1)$ such that

$$U^{m_\varnothing}(\hat{p}^*, \omega|\hat{p}^*) = U^{m_\omega}(\hat{p}^*, \omega|\hat{p}^*).$$

Given the simple expressive strategy with $\hat{p}^*$, $U^{m_\varnothing}(p, \omega|\hat{p}^*)$ is strictly increasing in p while $U^{m_\omega}(p, \omega|\hat{p})$ is constant in p. Hence, $U^{m_\varnothing}(p, \omega|\hat{p}^*) > (<) U^{m_\omega}(p, \omega|\hat{p}^*)$ if $p > (<) \hat{p}^*$. Thus, this is an equilibrium strategy.

Finally, suppose that the off-equilibrium beliefs are such that $\mu_\omega(\omega|m', n) = \pi(\omega)$ and $\mu_p(p|m', \omega) = \mu_p(p|m', n) = \psi(p)$ for any off-equilibrium message m'. Then no (p, ω) has any incentive to deviate.

A.6 *Proof of Fact 2*

First, I establish the following claim.

Claim 1. *Suppose $\lambda'' > \lambda'$. If there exists a unique simple expressive equilibrium outcome, the equilibrium cutoff $\hat{p}^*$ under λ' is higher than that under λ''.*

Recall that the type (p', ω) supervisor's expected payoff from $m_\varnothing$ is

$$U^{m_\varnothing}(p', \omega|\hat{p}, \lambda) = p'y(e(m_\varnothing, \omega|\hat{p})) + (1 - p')\lambda y(e(m_\varnothing, n|\hat{p})).$$

On the other hand, the supervisor's expected payoff from sending advice m_ω is

$$U^{m_\omega}(p', \omega|\hat{p}, \lambda) = y(e(m_\omega, \omega|\hat{p})).$$

Note that whenever there exists a unique simple expressive equilibrium outcome, the equilibrium cutoff type has to be unique. As I showed in the proof of Proposition 3, $\lim_{\hat{p}\to 1} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) > \lim_{\hat{p}\to 1} U^{m_\omega}(\hat{p}, \omega|\hat{p})$. Then we need to have $\lim_{\hat{p}\to 0} U^{m_\varnothing}(\hat{p}, \omega|\hat{p}) < \lim_{\hat{p}\to 0} U^{m_\omega}(\hat{p}, \omega|\hat{p})$ for the uniqueness of the cutoff. Clearly, $U^{m_\varnothing}(\hat{p}, \omega|\hat{p}, \lambda'') > U^{m_\varnothing}(\hat{p}, \omega|\hat{p}, \lambda')$, while $U^{m_\omega}(\hat{p}, \omega|\hat{p}, \lambda'') = U^{m_\omega}(\hat{p}, \omega|\hat{p}, \lambda')$. Then, since $U^{m_\varnothing}(\hat{p}, \omega|\hat{p}, \lambda)$ and $U^{m_\omega}(\hat{p}, \omega|\hat{p}, \lambda)$ are both strictly increasing in $\hat{p}$, whereas the uniqueness of the cutoff type implies that they intersect only once, the cutoff type $\hat{p}^*_{\lambda'}$ such that $U^{m_\varnothing}(\hat{p}^*_{\lambda'}, \omega|\hat{p}^*_{\lambda'}, \lambda') = U^{m_\omega}(\hat{p}^*_{\lambda'}, \omega|\hat{p}^*_{\lambda'}, \lambda')$ is always higher than $\hat{p}^*_{\lambda''}$ such that $U^{m_\varnothing}(\hat{p}^*_{\lambda''}, \omega|\hat{p}^*_{\lambda''}, \lambda') = U^{m_\omega}(\hat{p}^*_{\lambda''}, \omega|\hat{p}^*_{\lambda''}\lambda')$. This establishes Claim 1.

To prove Fact 2, observe that given a simple expressive strategy with cutoff $\hat{p}^*_\lambda$, the expected marginal disutility conditional on $m_\varnothing$ under λ is

$$\int_{p>\hat{p}^*_\lambda} c_1(e, p)\frac{\psi(p)}{1 - \Psi(\hat{p}^*_\lambda)}\, dp.$$

Since $c_{12}(e, p) < 0$, Claim 1 implies that the expected marginal disutility conditional on $m_\varnothing$ is higher under λ''. Then, since the competent agent's production function is always $y(e)$, we have $e(m_\varnothing, \omega; \lambda') > e(m_\varnothing, \omega; \lambda'')$.

On the other hand, given the simple expressive strategy with $\hat{p}^*_\lambda$, the expected disutility conditional on m_ω is

$$\int_{p<\hat{p}^*_\lambda} c_1(e, p)\frac{\psi(p)}{\Psi(\hat{p}^*_\lambda)}\, dp.$$

From Claim 1, the expected marginal disutility condition on m_ω is higher under λ''. Since m_ω reveals the true state, the agent's production function is always $y(e)$. Thus, $e(m_\omega, \omega; \lambda') > e(m_\omega, \omega; \lambda'')$ and $e(m_\omega, n; \lambda') > e(m_\omega, n; \lambda'')$.

A.7 *Proof of Fact 3*

Let $Y(m,h)$ be the output level in a simple expressive equilibrium conditional on m and h. First, since m_ω is sent by $p < \hat{p}^*$, whereas $m_\varnothing$ is sent by $p > \hat{p}^*$, $Y(m_\omega,\omega) < Y(m_\varnothing,\omega)$. Note that in the equilibrium, m_ω and $m_\varnothing$ have to be indifferent for the supervisor with $(\hat{p}^*,\omega)$. That is,

$$Y(m_\omega,\omega) = \hat{p}^* Y(m_\varnothing,\omega) + (1-\hat{p}^*)Y(m_\varnothing,n).$$

It follows that $Y(m_\varnothing,n) < Y(m_\omega,\omega) < Y(m_\varnothing,\omega)$.

Moreover, $\mu_p(p|m_\varnothing,\omega)$ first-order-stochastically dominates $\psi(p)$, while $\psi(p)$ first-order-stochastically dominates $\mu_p(p|m_\omega,\omega)$. Thus, from $c_{12}(e,p) < 0$, $e(m_\varnothing,\omega) > \tilde{e} > e(m_\omega,\omega)$. Hence, $Y(m_\omega,\omega) < y(\tilde{e}) < Y(m_\varnothing,\omega)$. This implies that there exists $\tilde{p} \in (\hat{p}^*,1)$ such that

$$\tilde{p}Y(m_\varnothing,\omega) + (1-\tilde{p})Y(m_\varnothing,n) = y(\tilde{e}).$$

Note that $y(\tilde{e})$ is the supervisor's payoff in fully directive equilibrium. Then, if $p > \tilde{p}$, the supervisor prefers the simple expressive equilibrium to a fully directive equilibrium while $p < \tilde{p}$ prefers the opposite.

A.8 *Proof of Proposition 4*

The basic idea of the proof is analogous to that in the proof of Proposition 3. Suppose the supervisor uses a strategy in Proposition 4. Clearly, since $\lambda < (1-\delta)\pi(\omega_{\max}) + \delta$, $a(m_{\max},n) = a_{\omega_{\max}}$ given any cutoff $\hat{p}$. Let $e(m,h|\hat{p})$ be the optimal effort level given the strategy with cutoff $\hat{p}$. Note that the type (p,ω) supervisor's expected payoff from $m_{\max}$ is

$$U^{m_{\max}}(p,\omega|\hat{p}) = py(e(m_{\max},\omega|\hat{p})) + (1-p)\delta y(e(m_{\max},n|\hat{p}))$$

if $\omega \neq \omega_{\max}$. On the other hand, since m_ω always reveals the true state, $e(m_\omega,\omega|\hat{p}) = e(m_\omega,n|\hat{p})$. Thus, the supervisor's expected payoff from m_ω is

$$U^{m_\omega}(p,\omega|\hat{p}) = y(e(m_\omega,\omega|\hat{p})).$$

For the competent agent at $\omega \neq \omega_{\max}$, his belief about p conditional on $m_{\max}$ is the same as that conditional on $m_\varnothing$ given a simple expressive strategy with cutoff $\hat{p}$. On the other hand, when the incompetent agent receives $m_{\max}$, he cannot rule out the possibility of $\omega = \omega_{\max}$. Note that this makes his belief about p different from that given a simple expressive strategy. However, as $\hat{p}$ goes to 1, the supervisor's payoff from $m_{\max}$ is the same as that from $m_\varnothing$ in the proof of Proposition 3. Note that as $\hat{p}$ goes to 1, the only p who uses the message in $\omega \neq \omega_{\max}$ becomes $p = 1$. That is, $\mu_p(p=1|m_{\max},\omega)$ converges to 1 if $\omega \neq \omega_{\max}$. It implies that the competent agent's effort level converges to $e_1 = \arg\max\{y(e) - c(e,1)\}$ as $\hat{p}$ goes to 1. Moreover, as $\hat{p}$ goes to 1, the supervisor with $\hat{p}$ believes that the agent is competent for sure. Thus, $\lim_{\hat{p}\to 1} U^{m_{\max}}(\hat{p},\omega|\hat{p}) = y(e_1)$ as in

the case of $m_\varnothing$ in the proof of Proposition 3. Thus,

$$\lim_{\hat{p}\to 1} U^{m_{\max}}(\hat{p}, \omega|\hat{p}) > \lim_{\hat{p}\to 1} U^{m_\omega}(\hat{p}, \omega|\hat{p})$$

for $\omega \neq \omega_{\max}$.

If $\hat{p} = 0$, the strategy is just a pooling strategy and $m_{\max}$ induces $a_{\omega_{\max}}$. Thus, $U^{m_{\max}}(\hat{p}, \omega|\hat{p}) = \delta y(e(m_{\max}, n|0))$ if $\hat{p} = 0$ and $\omega \neq \omega_{\max}$. Since $m_{\max}$ induces a wrong technology, $e(m_{\max}, n|0) < \tilde{e}$. Moreover, since $\mu_p(p = 0|m_\omega, h)$ converges to 1 as $\hat{p}$ goes to 0, $\lim_{\hat{p}\to 0} U^{m_\omega}(\hat{p}, \omega|\hat{p}) = y(e_0)$ if $\omega \neq \omega_{\max}$. Note that $e(m_{\max}, n|0) \leq \tilde{e}$. Therefore, if $\delta y(\tilde{e}) < y(e_0)$, then

$$\lim_{\hat{p}\to 0} U^{m_{\max}}(\hat{p}, \omega|\hat{p}) < \lim_{\hat{p}\to 0} U^{m_\omega}(\hat{p}, \omega|\hat{p}).$$

Note that since $U^{m_{\max}}(\hat{p}, \omega|\hat{p}) - U^{m_\omega}(\hat{p}, \omega|\hat{p})$ is continuous in $\hat{p} \in (0, 1)$, there exists at least one $\hat{p}^* \in (0, 1)$ such that

$$U^{m_{\max}}(\hat{p}^*, \omega|\hat{p}^*) = U^{m_\omega}(\hat{p}^*, \omega|\hat{p}^*).$$

To see that the strategy with $\hat{p}^*$ is an equilibrium strategy, note that $U^{m_{\max}}(p, \omega|\hat{p}^*)$ is strictly increasing in p, while $U^{m_\omega}(p, \omega|\hat{p})$ is constant in p. Hence, $U^{m_{\max}}(p, \omega|\hat{p}^*) > (<) U^{m_\omega}(p, \omega|\hat{p}^*)$ if $p > (<)\hat{p}^*$ for $\omega \neq \omega_{\max}$.

Now suppose that the off-equilibrium beliefs are such that $\mu_\omega(\omega|m', n) = \pi(\omega)$ and $\mu_p(p|m', \omega) = \mu_p(p|m', n) = \psi(p)$ for any off-equilibrium message m'. Then $a(m', n) = a_{\omega_{\max}}$. Since $\mu_p(p|m_{\max}, h)$ first-order-stochastically dominates $\psi(p)$, $U^{m_{\max}}(p, \omega|\hat{p}^*) \geq U^{m'}(p, \omega|\hat{p}^*)$ for any (p, ω) who sends $m_{\max}$ in the strategy. Moreover, from the equilibrium condition, $U^{m_{\max}}(\hat{p}^*, \omega|\hat{p}^*) = U^{m_\omega}(p, \omega|\hat{p}^*)$ for any p if $\omega \neq \omega_{\max}$. Then, since $U^{m_{\max}}(\hat{p}^*, \omega|\hat{p}^*) \geq U^{m'}(p, \omega|\hat{p}^*)$, any (p, ω) who sends m_ω in the strategy has no incentive to send m'.

A.9 *Proof of Proposition 5*

The supervisor's expected payoff in fully directive equilibrium is $y(\tilde{e})$. When the safe choice is the incompetent agent's optimal choice given the prior, the expected payoff from $D(\hat{p}) = (\hat{p}, 1] \times \Omega$ is

$$U^{D(\hat{p})}(p, \omega) = py(e(D(\hat{p}), \omega)) + (1 - p)\lambda y(e(D(\hat{p}), n)),$$

where $e(D, h)$ is the agent's best response to D given h. Clearly, $y(\tilde{e}) > U^{D(\hat{p})}(\hat{p}, \omega)$ for sufficiently small $\hat{p}$ whereas $U^{D(\hat{p})}(\hat{p}, \omega) > y(\tilde{e})$ for sufficiently large $\hat{p}$. Then, since $U^{D(\hat{p})}(p, \omega)$ is strictly increasing and continuous in $\hat{p}$, there exists $\hat{p}^{**} \in (\hat{p}^*, 1)$ such that $U^{D(\hat{p}^{**})}(\hat{p}^{**}, \omega) = y(\tilde{e})$. Since $U^{D(\hat{p})}(p, \omega)$ is strictly increasing in p, $U^D(p, \omega) > (<) y(\tilde{e})$ if $(p, \omega) \in (\notin) D$. Thus, $D = (\hat{p}^{**}, 1] \times \Omega$ is a self-signaling set.

A.10 *Proof of Proposition 6*

Suppose there exists a simple expressive equilibrium. Let $D = Q \times Z \subset P \times \Omega$ be a self-signaling set. Moreover, let $a(D, h)$ be the induced technology choice and let $e(D, h)$ be the induced effort level induced by D given h. That is,

$$e(D,h) = \arg\max_{e \in E} \left\{ \sum_{\omega} f(e, a(D,h), \omega) \mu_{\omega}(\omega | D, h) - \int_{p} c(e,p) \mu_{p}(p|D,h)\, dp \right\}.$$

Let $U^D(p, \omega)$ denote the supervisor (p, ω)'s expected payoff from D. First, I establish the following lemma.

LEMMA 4. *If a self-signaling set exists for a simple expressive equilibrium, then* $D = [0, q) \times \{\omega\}$, *where* $q > \hat{p}^*$.

To prove the lemma, consider the following three cases.

Case 1: $|Z| = 1$. Let $\{\omega'\} = Z$.

Step 1. If $p' \in Q$ and $p' < \hat{p}^*$, $[0, \hat{p}^*) \subset Q$. Suppose not. Then there exists $p'' < \hat{p}^*$ such that $p'' \notin Q$. Since D is a self-signaling set, $U^D(p', \omega') > U^{m_{\omega'}}(p', \omega')$. Since $a(D, h) = a_{\omega'}$ for $h = n, \omega'$, $U^D(p', \omega') = U^D(p'', \omega')$, whereas $U^{m_{\omega'}}(p', \omega') = U^{m_{\omega'}}(p'', \omega') = y(e(m_{\omega'}, \omega'))$. Therefore, $U^D(p'', \omega') > U^{m_{\omega}}(p'', \omega')$, a contradiction.

Step 2. If $p'' \in Q$ and $p'' > \hat{p}^*$, then $(\hat{p}^*, p'') \subset Q$. Suppose not. Then there exists $p' \in (\hat{p}^*, p'')$ such that $p' \notin Q$. Since D is a self-signaling set, $U^D(p'', \omega') > U^{m_{\varnothing}}(p'', \omega')$. Since $a(D, h) = a_{\omega'}$ for $h = n, \omega'$, $U^D(p', \omega') = U^D(p'', \omega')$. On the other hand, since $U^{m_{\varnothing}}(p, \omega)$ is strictly increasing in p, $U^{m_{\varnothing}}(p'', \omega) > U^{m_{\varnothing}}(p', \omega)$. Thus, $U^D(p', \omega') > U^{m_{\varnothing}}(p', \omega')$, a contradiction.

Step 3: $Q \neq (\hat{p}^*, p'')$ for any $p'' \in (\hat{p}^*, 1]$. Suppose not. Then, for some $p'' \in (\hat{p}^*, 1]$, $U^{\varnothing}(p, \omega') < U^D(p, \omega')$ for any $p \in (\hat{p}^*, p'')$. Moreover, since $a(D, h) = a_{\omega'}$ for any h, $U^D(p, \omega')$ is constant in p. But since $U^{m_{\omega}}(p, \omega') = U^{\varnothing}(\hat{p}^*, \omega')$ for any $p \leq \hat{p}^*$, $U^{m_{\omega}}(p, \omega') < U^D(p, \omega')$ for any $p \leq \hat{p}^*$, a contradiction.

From Step 1–3, whenever $Z = \{\omega'\}$, $D = [0, q) \times \{\omega'\}$, where $q \geq \hat{p}^*$.

Now turning to self-signaling sets with $|Z| > 1$, consider the following two cases.

Case 2: $|Z| > 1$ and $a(D, n) = a_0$. In this case, the type (p, ω) supervisor's expected payoff given D is $U^D(p, \omega) = py(e(D, \omega)) + (1 - p)\lambda y(e(D, n))$. Since the agent's belief about p given Q is independent of h, $e(D, \omega) > e(D, n)$. Thus, $U^D(p, \omega)$ is strictly increasing in p.

Step 1. If $p' \in Q$ and $p' > \hat{p}^*$, then $(\hat{p}^*, 1] \subset Q$. Suppose $p'' \notin Q$ for some $p'' > \hat{p}^*$. Note that $a(D, \omega) = a(m_{\varnothing}, \omega) = a_{\omega}$ and $a(D, n) = a(m_{\varnothing}, n) = a_0$. Thus, if $U^D(p', \omega) > U^{m_{\varnothing}}(p', \omega)$, then $e(D, h) > e(m_{\varnothing}, h)$ for $h = n, \omega$. Then we also have $U^D(p'', \omega) > U^{m_{\varnothing}}(p'', \omega)$, a contradiction.

Step 2: $Q \neq (\hat{p}^*, 1]$. Note that $P_{\sigma}(m_{\varnothing}, \omega) = (\hat{p}^*, 1]$ in simple expressive equilibrium. Thus, if $Q = (\hat{p}^*, 1]$, then $e(D, h) = e(m_{\varnothing}, h)$ for $h = n, \omega$ and thus $U^D(p, \omega) = U^{m_{\varnothing}}(p, \omega)$ for any $p \in Q$, a contradiction.

Step 3. If $p' \in Q$ and $p' < p'' < \hat{p}^*$, then $p'' \in Q$. Suppose not. Then $p' \in Q$ but $p'' \notin Q$. Note that $U^D(p, \omega)$ is strictly increasing in p, while $U^{m_{\omega}}(p, \omega)$ is constant in p. Hence, $U^D(p'', \omega) > U^{m_{\omega}}(p'', \omega)$ whenever $U^D(p', \omega) > U^{m_{\omega}}(p', \omega)$, a contradiction.

From Steps 1 and 2, if $|Z| > 1$, $a(D, n) = a_0$ and $\sup Q > \hat{p}^*$, then $Q \supsetneq (\hat{p}^*, 1]$. Then Step 3 implies that the only possibility is $Q = (\tilde{p}, 1]$, where $\tilde{p} < \hat{p}^*$.

Step 4: $Q \neq (\tilde{p}, 1]$, where $\tilde{p} < \hat{p}^*$. Suppose $Q = (\tilde{p}, 1]$. Then, since $\tilde{p} < \hat{p}^*$, $\mu_p(p|m_\varnothing, h)$ first-order-stochastically dominates $\mu_p(p|D, h)$ for $h = \omega, n$. Then, since D and $m_\varnothing$ induce the same technology choice given $h = n, \omega$, $e(D, h) < e(m_\varnothing, h)$ for $h = n, \omega$. But then $U^D(p, \omega) < U^{m_\varnothing}(p, \omega)$ for $p \in Q$, a contradiction.

Step 4 implies that $\sup Q \leq \hat{p}^*$. From Step 3, if $|Z| > 1$, $a(D, n) = a_0$, and $\sup Q \leq \hat{p}^*$, then the only possibility is $Q = (\tilde{p}, \hat{p}^*]$ or $(\tilde{p}, \hat{p}^*)$, where $\tilde{p} \in [0, \hat{p}^*)$.

Step 5. There is no self-signaling set such that $|Z| > 1$ and $a(D, n) = a_0$. Let $U^{D(\tilde{p})}(p, \omega)$ be the expected payoff of (p, ω) from D when $Q = (\tilde{p}, \hat{p}^*]$ and $\omega \in Z$. That is,

$$U^{D(\tilde{p})}(p, \omega) = py(e(D(\tilde{p}), \omega)) + (1 - p)\lambda y(e(D(\tilde{p}), n)).$$

Note that $\lim_{\tilde{p} \to \hat{p}^*} E_p[c(e, p)|D(\tilde{p})] = c(e, \hat{p}^*)$. Then $\lim_{\tilde{p} \to \hat{p}^*} e(D(\tilde{p}), h) < e(m_\varnothing, h)$ for $h = n, \omega$. Hence, $\lim_{\tilde{p} \to \hat{p}^*} U^{D(\tilde{p})}(\tilde{p}, \omega) < U^{m_\varnothing}(\hat{p}^*, \omega)$. Note that from the equilibrium condition, $U^{m_\varnothing}(\hat{p}^*, \omega) = U^{m_\omega}(\hat{p}^*, \omega)$. Since $U^{m_\omega}(p, \omega) = y(e(m_\omega, \omega))$ for any $p \in [0, \hat{p}^*]$, $\lim_{\tilde{p} \to \hat{p}^*} U^{D(\tilde{p})}(\tilde{p}, \omega) < U^{m_\omega}(p, \omega)$. Then, since $U^{D(\tilde{p})}(\tilde{p}, \omega)$ is strictly increasing in $\tilde{p}$, $U^{D(\tilde{p})}(\tilde{p}, \omega) < U^{m_\omega}(p, \omega)$ for any $\tilde{p} < \hat{p}^*$, a contradiction.

Case 3: $|Z| > 1$ and $a(D, n) \neq a_0$. Suppose there exists a self-signaling set with $|Z| > 1$ and $a(D, n) \neq a_0$. If $Z = \Omega$, then $a(D, n) = a_0$ since there exists a simple expressive equilibrium. Thus, consider the case where $Z \subsetneq \Omega$ and let $\omega(D)$ be ω such that $a(D, n) = a_\omega$. Then $\omega(D) \in Z$. Since $|Z| > 1$ and $Z \subsetneq \Omega$, we can always find $\omega' \in Z$ such that $\omega' \neq \omega(D)$ and $\omega'' \notin Z$. Then, by definition, for any $p' \in Q$, $U^D(p', \omega') > U^{\sigma(p', \omega')}(p', \omega')$. On the other hand, $U^{\sigma(p', \omega'')}(p', \omega'') = U^{\sigma(p', \omega')}(p', \omega')$ in any simple expressive equilibrium. Then, since $U^D(p', \omega') = U^D(p', \omega'')$, $U^D(p', \omega'') > U^{\sigma(p', \omega'')}(p', \omega'')$, a contradiction.

From Cases 2 and 3, there is no self-signaling set with $|Z| > 1$.

Now, we are ready to prove Proposition 6.

From Lemma 4, when a credible neologism exists in a simple expressive equilibrium, it always takes the form of $[0, q) \times \{\omega\}$, where $q > \hat{p}^*$. Let $D(q) = [0, q) \times \{\omega\}$ and $e(q) = \arg\max_{e \in E}\{y(e) - \int_{p<q} c(e, p; \beta)\psi(p)/\Psi(q)\, dp\}$. Moreover, let $U^{D(q)}(p, \omega)$ be the type (p, ω) supervisor's payoff when the agent optimally responds to $D(q)$, that is, $U^{D(q)}(q, \omega) = y(e(q))$.

From the indifference condition of simple expressive equilibrium, $U^{D(\hat{p}^*)}(\hat{p}^*, \omega) = U^{m_\omega}(\hat{p}^*, \omega)$. Thus, $U^{D(q)}(q, \omega)$ can be written as

$$U^{D(q)}(q, \omega) = U^{m_\omega}(\hat{p}^*, \omega) + \int_{\hat{p}^*}^{q} y'(e(q'))e'(q')\, dq'.$$

Since $e(q')$ is the optimal effort level given $D(q')$, it satisfies the first-order condition:

$$y'(e(q')) - \frac{1}{\Psi(q')} \int_0^{q'} c_1(e(q'), p; \beta)\psi(p)\, dp = 0.$$

Then, by the implicit function theorem,

$$e'(q') = -\frac{\frac{\psi(q')}{\Psi(q')}[\frac{1}{\Psi(q')}\int_0^{q'} c_1(e(q'), p; \beta)\psi(p)\,dp - c_1(e(q'), q'; \beta)]}{y''(e(q')) - \frac{1}{\Psi(q')}\int_0^{q'} c_{11}(e(q'), p; \beta)\psi(p)\,dp}. \tag{1}$$

Since $y''(e(q')) \leq 0$ and $c_{11}(e(q'), p; \beta) > 0$ for any p, the denominator of the right hand side of (1) is strictly negative. On the other hand, since $c_{12}(e, p; \beta) < 0$, the numerator of the right hand side of (1) is strictly positive. Thus, $e'(q') > 0$.

Note that

$$\frac{1}{\Psi(q')}\int_0^{q'} c_1(e(q'), p; \beta)\psi(p)\,dp - c_1(e(q'), q'; \beta) < c_1(e(q'), 0; \beta) - c_1(e(q'), 1; \beta).$$

Since $c_1(e, 0; \beta) - c_1(e, 1; \beta) = \beta[b_1(e, 1) - b_1(e, 0)]$, the numerator of the right hand side of (1) converges to 0 as β goes to 0. Let $e_l = \arg\max_{e\in E}\{y(e) - C(e)\}$. Since $e_l < e(q')$ for any $q' \geq \hat{p}^*$ and $b_{11}(e, p) \leq 0$, the denominator of the right hand side of (1) is strictly smaller than $-C''(e_l) < 0$ for any $\beta \in (0, 1]$. Hence, for any $\epsilon > 0$, $e'(q') < \epsilon$ if β is sufficiently small.

From the equilibrium condition, $U^{m_\varnothing}(q, \omega)$ can be written as

$$U^{m_\varnothing}(q, \omega) = U^{m_\omega}(\hat{p}^*, \omega) + \int_{\hat{p}^*}^{q} U_1^{m_\varnothing}(p, \omega)\,dp.$$

Since $e(m_\varnothing, \omega) > e_l$,

$$U_1^{m_\varnothing}(p, \omega) = y(e(m_\varnothing, \omega)) - \lambda y(e(m_\varnothing, n)) > (1 - \lambda)y(e_l) > 0.$$

Thus, if β is sufficiently small, $U^{D(q)}(q, \omega) \leq U^{m_\varnothing}(q, \omega)$ for any $q' \in [\hat{p}^*, 1]$. That is, there is no self-signaling set.

A.11 *Proof of Fact 4*

In simple expressive equilibrium, the equilibrium cutoff $\hat{p}$ makes m_ω and $m_\varnothing$ indifferent to the supervisor. In this setting, the indifference condition can be simplified to the quadratic equation

$$\hat{p}^2 + \frac{2\alpha(1+\lambda)}{\beta}\hat{p} - \frac{2\alpha(1+\lambda)}{\beta} + \frac{\lambda}{1-\lambda} = 0.$$

Note that the right hand side has the minimum value at $\hat{p} = -\alpha(1+\lambda)/\beta < 0$ whenever $\alpha, \beta, \lambda > 0$. Then, since there is at most one solution with a positive value, the equilibrium cutoff is unique whenever it exists. Since there exists a unique cutoff type that solves the equation, the existence of a simple expressive equilibrium requires $\lambda < y(e_0)/y(e_\lambda)$. Since $e_\lambda = \alpha\lambda + \beta/2$ and $e_0 = \alpha$, $\lambda < y(e_0)/y(e_\lambda)$ is simplified to $\lambda\beta/(2(1-\lambda^2)) < \alpha$.

Turning to neologism proofness, if $(\partial/\partial q)U^{D(q)}(p, \omega) \leq U_1^{m_\varnothing}(p, \omega)$ for any $q \geq \hat{p}^*$, then $U^{D(q)}(p, \omega) \leq U^{m_\varnothing}(p, \omega)$ for any $p \geq \hat{p}^*$ and the simple expressive equilibrium is

neologism-proof. Observe that

$$U_1^{m_\varnothing}(p,\omega) = y(e(m_\varnothing,\omega)) - \lambda y(e(m_\varnothing,n)) > (1-\lambda)y(\tilde{e}) = \alpha^2(1-\lambda) + \frac{\alpha}{2}(1-\lambda)\beta.$$

Moreover, since $e(q) = \alpha + \beta q/2$, $U^{D(q)}(p,\omega) = \alpha(\alpha + \beta q/2)$. Thus,

$$\frac{\partial}{\partial q}U^{D(q)}(p,\omega) = \frac{\alpha\beta}{2}.$$

By inspection, if $U_q^{D(q)}(p,\omega) \leq \alpha^2(1-\lambda) + \alpha/2(1-\lambda)\beta$, then $\lambda\beta/(2(1-\lambda)) \leq \alpha$. Since $U_1^{m_\varnothing}(p,\omega) > \alpha^2(1-\lambda) + \alpha/2(1-\lambda)\beta$, $\lambda\beta/(2(1-\lambda)) \leq \alpha$ implies $U_q^{D(q)}(p,\omega) < U_1^{m_\varnothing}(p,\omega)$. Note that since $\lambda \in (0,1)$, $\lambda\beta/(2(1-\lambda^2)) > \lambda\beta/(2(1-\lambda))$. Thus, whenever there exists a simple expressive equilibrium, it satisfies the condition of neologism proofness.

A.12 *Neologism-proof equilibrium and multiple simple expressive equilibria*

FACT 6. *If there are multiple simple expressive equilibrium outcomes, only those with the highest cutoff type can be neologism-proof.*

PROOF. As in the proof of Proposition 3, let $U^m(p,\omega|p)$ be the supervisor with (p,ω)'s expected payoff from m given the simple expressive strategy with cutoff p. Let $\hat{p}^*$ be the highest cutoff type of simple expressive equilibria.

Suppose there exists an equilibrium cutoff $\hat{p}^{**}$ such that $\hat{p}^{**} < \hat{p}^*$. From the equilibrium condition, $U^{m_\varnothing}(\hat{p}^{**},\omega|\hat{p}^{**}) = U^{m_\varnothing}(\hat{p}^{**},\omega|\hat{p}^{**})$ and $U^{m_\varnothing}(\hat{p}^*,\omega|\hat{p}^*) = U^{m_\varnothing}(\hat{p}^*,\omega|\hat{p}^*)$. Then, since $\lim_{p\to 1} U^{m_\omega}(p,\omega|p) < \lim_{p\to 1} U^{m_\varnothing}(p,\omega|p)$ while $U^{m_\omega}(p,\omega|p)$ and $U^{m_\varnothing}(p,\omega|p)$ are continuous in p, there exists $p' \in [\hat{p}^{**},\hat{p}^*]$ such that $U^{m_\omega}(p',\omega|p') > U^{m_\varnothing}(p',\omega|p')$. On the other hand, since $U^{m_\varnothing}(p,\omega|p) > U^{m_\varnothing}(p,\omega|\hat{p}^{**})$ for any $p > \hat{p}^{**}$, $U^{m_\omega}(p',\omega|p') > U^{m_\varnothing}(p',\omega|\hat{p}^{**})$. Note that $U^{m_\omega}(1,\omega|1) < U^{m_\varnothing}(1,\omega|\hat{p}^{**})$. Then, since $U^{m_\omega}(p,\omega|p)$ and $U^{m_\varnothing}(p,\omega|\hat{p}^{**})$ are continuous in p, there exists $p'' \in (p',1)$ such that $U^{m_\omega}(p'',\omega|p'') = U^{m_\varnothing}(p'',\omega|\hat{p}^{**})$.

Since $U^{m_\varnothing}(p,\omega|\hat{p}^{**})$ is strictly increasing in p whereas $U^{m_\omega}(p,\omega|p'')$ is constant in p, $U^{m_\omega}(p,\omega|p'') > U^{m_\varnothing}(p,\omega|\hat{p}^{**})$ for any $p < p''$ and $U^{m_\omega}(p,\omega|p'') < U^{m_\varnothing}(p,\omega|\hat{p}^{**})$ if $p > p''$. Note that $U^{m_\omega}(p,\omega|p'')$ is the same as the supervisor's payoff from neologism $[0,p'') \times \{\omega\}$. Hence, $[0,p'') \times \{\omega\}$ is a self-signaling set. □

A.13 *A related signaling game and the D1 equilibrium*

Suppose $M = A$. Assume that the incompetent's choice function is restricted so that $a(m,n) = m$. This transforms the current game into a costly signaling game with single type p. In the fully directive equilibrium, the supervisor's strategy is $\sigma(p,\omega) = a_\omega$. On the other hand, in the simple expressive equilibrium, the supervisor's strategy is $\sigma(p,\omega) = a_\omega$ if $p < \hat{p}$; $\sigma(p,\omega) = a_0$ if $p > \hat{p}$.

The simple expressive equilibria always passes the D1 criterion whenever it exists. Thus, the question is whether the fully directive equilibrium passes the D1 criterion.

FACT 7. *The fully directive equilibrium fails the D1 criterion whenever the simple expressive equilibrium exists.*

PROOF. Let

$$e_\omega(\mu_p) = \arg\max_{e\in E}\left\{y(e) - \int_p c(e,p)\mu_p(p)\,dp\right\}$$

$$e_n(\mu_p) = \arg\max_{e\in E}\left\{\lambda y(e) - \int_p c(e,p)\mu_p(p)\,dp\right\}.$$

Then define the best response efforts to m given μ_p whose support is Q:

$$BR(Q,m) = \bigcup_{\{\mu_p:\mathrm{supp}(\mu_p)=Q\}} \{(e_\omega(\mu_p), e_n(\mu_p))\}.$$

By the restriction on $a(m,n)$, $\sigma(p,\omega) = a_\omega$ for all (p,ω) in the fully directive equilibrium. To test whether the fully directive equilibrium passes D1, suppose that the supervisor sends off-equilibrium message $m = a_0$. Then define

$$\mathcal{D}_p(\omega, a_0) := \{(e_\omega, e_n) \in BR(P, a_0) : y(\tilde{e}) < py(e_\omega) + (1-p)\lambda y(e_n)\},$$

which is the set of best response efforts to a_0 that makes the supervisor's payoff strictly higher than that in the fully directive equilibrium.

Since $(e_\omega, e_n) \in BR(P, a_0)$, $e_\omega > e_n$ if $(e_\omega, e_n) \in \mathcal{D}_p(\omega, a_0)$. Hence, if $(e_\omega, e_n) \in \mathcal{D}_p(\omega, a_0)$, then $y(\tilde{e}) < p'y(e_\omega) + (1-p')\lambda y(e_n)$ for any $p' > p$, that is, $(e_\omega, e_n) \in \mathcal{D}_{p'}(\omega, a_0)$.

Now define

$$\mathcal{D}^0_p(\omega, a_0) := \{(e_\omega, e_n) \in BR(P, a_0) : y(\tilde{e}) = py(e_\omega) + (1-p)\lambda y(e_n)\}.$$

If $(e_\omega, e_n) \in \mathcal{D}^0_p(\omega, a_0)$, then $e_\omega > e_n$ since $(e_\omega, e_n) \in BR(P, a_0)$. Thus, if $y(\tilde{e}) = py(e_\omega) + (1-p)\lambda y(e_n)$, then $y(\tilde{e}) < p'y(e_\omega) + (1-p')\lambda y(e_n)$ for any $p' > p$, that is, $(e_\omega, e_n) \in \mathcal{D}_{p'}(\omega, a_0)$. Therefore,

$$\mathcal{D}_p(\omega, a_0) \cup \mathcal{D}^0_p(\omega, a_0) \subseteq \mathcal{D}_{p'}(\omega, a_0).$$

Then the D1 criterion requires that $\mu_p(p|a_0) = 0$ for any $p < 1$. Under the off-equilibrium belief, the supervisor with a sufficiently high p has an incentive to send $m = a_0$; that is, the D1 criterion eliminates the fully directive equilibrium. □

REFERENCES

Austin, John L. (1962), *How to Do Things With Words*. The William James Lectures Delivered at Harvard University in 1955. Oxford, Clarendon Press, London. [178]

Battigalli, Pierpaolo and Martin Dufwenberg (2009), "Dynamic psychological games." *Journal of Economic Theory*, 144, 1–35. [180]

Bénabou, Roland and Jean Tirole (2003), "Intrinsic and extrinsic motivation." *Review of Economic Studies*, 70, 489–520. [179]

Bernheim, B. Douglas (1994), "A theory of conformity." *Journal of Political Economy*, 102, 841–877. [180]

Blume, Andreas and Oliver Board (2013), "Language barriers." *Econometrica*, 81, 781–812. [178, 179, 194]

Brown, Penelope and Stephen C. Levinson (1987), *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge. [178]

Chakraborty, Archishman and Rick Harbaugh (2010), "Persuasion by cheap talk." *American Economic Review*, 100, 2361–2382. [179]

Chen, Ying (2011), "Perturbed communication games with honest senders and naive receivers." *Journal of Economic Theory*, 146, 401–424. [194]

Crawford, Vincent P. and Joel Sobel (1982), "Strategic information transmission." *Econometrica*, 50, 1431–1451. [178]

Enzle, Michael E. and Sharon C. Anderson (1993), "Surveillant intentions and intrinsic motivation." *Journal of Personality and Social Psychology*, 64, 257–266. [180]

Farrell, Joseph (1993), "Meaning and credibility in cheap-talk games." *Games and Economic Behavior*, 5, 514–531. [189]

Farrell, Joseph and Robert Gibbons (1989), "Cheap talk with two audiences." *American Economic Review*, 79, 1214–1223. [179]

Geanakoplos, John, David Pearce, and Ennio Stacchetti (1989), "Psychological games and sequential rationality." *Games and Economic Behavior*, 1, 60–79. [180]

Glazer, Jacob and Ariel Rubinstein (2001), "Debates and decisions: On a rationale of argumentation rules." *Games and Economic Behavior*, 36, 158–173. [177]

Glazer, Jacob and Ariel Rubinstein (2006), "A study in the pragmatics of persuasion: A game theoretical approach." *Theoretical Economics*, 1, 395–410. [177]

Grice, Paul (1989), *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts. [177]

Köszegi, Botond (2006), "Emotional agency." *Quarterly Journal of Economics*, 121, 121–155. [180]

Morgan, John and Phillip Stocken (2003), "An analysis of stock recommendations." *Rand Journal of Economics*, 34, 183–203. [178]

Morris, Stephen (2001), "Political correctness." *Journal of Political Economy*, 109, 231–265. [178]

Rubinstein, Ariel (2000), *Economics and Language*. Cambridge University Press, Cambridge. [177]

Searle, John R. (1975), "A taxonomy of illocutionary acts." In *Language, Mind, and Knowledge* (Keith Gunderson, ed.). University of Minnesota Press, Minneapolis. [178]

Co-editor Johannes Hörner handled this manuscript.

Manuscript received 10 May, 2014; final version accepted 24 January, 2016; available online 27 January, 2016.