

Bien opérer Perun

Un manuel pour les personnes qui répondent aux questions de Perun — et pour quiconque veut le devenir.

Zerava Technologies Inc. · Sous licence Creative Commons Attribution - Partage dans les mêmes conditions 4.0 International (CC BY-SA 4.0) : vous pouvez le copier, le partager et l'adapter, y compris à des fins commerciales, à condition d'en citer la source et de partager vos adaptations sous la même licence. Traduction de « Operating Perun well ».

Avant de commencer

Ce livre ne demande aucune connaissance préalable en IA ni en sécurité. Il porte sur **la manière de réfléchir quand vous répondez à une question qui compte** : une question sur une seule action, posée par un système qui a arrêté quelque chose et qui veut qu'une personne décide.

Vous pouvez le lire avant même de voir Perun. Quand vous recevrez un accès, vous ferez d'abord un court **exercice** : une série d'appels retenus, de déclarations et de questions, conçus pour l'exercice et tenus à l'écart du vrai travail. L'exercice est mesuré, et vous le saurez ; il mesure les mêmes choses que ce livre enseigne. Rien dans l'exercice ne vous interroge sur vous-même. Il regarde seulement comment vous travaillez.

Chaque chapitre contient des exemples, une règle et des exercices. Les réponses se trouvent à la fin du livre.

1. Le travail de l'opérateur

Un exemple. Perun retient un appel : l'agent d'IA de l'entreprise veut supprimer un dossier de vieilles factures. Quelqu'un doit dire oui ou non. Ce quelqu'un, c'est l'opérateur.

La mission. Perun tient une mémoire et un jugement appelés **SHARED_LIB**. Sa mission est de laisser travailler les agents d'IA de votre entreprise tout en arrêtant ce qui ne doit pas arriver. La plupart du temps, il décide seul. Quand il ne le peut pas — l'action est irréversible et la demande vient de quelqu'un qu'il ne connaît pas encore, ou la demande cache ce qu'elle veut dire —, il demande à un humain. **Votre mission d'opérateur est de faire en sorte que cette mission**

continue d'avancer : répondre à ce qui est demandé, à temps, avec une raison, sur la chose précise qui a été demandée.

L'équilibre, pas la perfection. Perun mesure tout ce qui est vivant en lui avec un nombre appelé **theta**, normalement à **30**, avec le droit de s'écarter de **cinq dans un sens ou dans l'autre**. Les opérateurs en ont un aussi. Vous dérivez dans un sens si vous laissez trop passer (approuver sans lire, décider trop vite, poser des questions vagues) ; vous dérivez dans l'autre sens si vous freinez trop (refuser presque tout, laisser des questions en attente, annuler vos propres réponses). Un bon opérateur oscille autour de 30 et y revient — comme un bon conducteur qui corrige sans cesse un peu le volant et ne donne jamais de coup brusque.

La règle. Vous n'êtes pas là pour avoir raison sur tout. Vous êtes là pour répondre à une question concrète à la fois, avec soin, pour que le travail continue.

Exercices 1.1 En une phrase, quelle est la mission d'un opérateur ? 1.2 Nommez deux façons dont un opérateur peut dériver vers *laisser trop passer*, et deux vers *trop freiner*.

2. Concret et abstrait

Un exemple. Deux opérateurs décrivent le même appel retenu.

- Le premier : « *L'agent veut envoyer le fichier Q3-salaries.xlsx à pay@unknown-domain.biz.* »
- Le second : « *L'agent veut faire quelque chose avec des données sensibles.* »

La première description est **concrète** : vous pouvez en désigner chaque partie — ce fichier, cette adresse, cette action. La seconde est **abstraite** : elle nomme une catégorie. Les deux sont vraies. Seule la première permet de décider.

Pourquoi c'est important. Perun pose des questions sur des actions concrètes. Ses déclarations sont concrètes elles aussi : *cet outil peut supprimer des fichiers ; il ne peut toucher que le dossier /work/invoices*. Un opérateur qui transforme des questions concrètes en questions abstraites — « *est-ce sensible ?* », « *cette source est-elle correcte en général ?* » — perd ce sur quoi portait la question et répond à autre chose.

Quand l'abstrait aide. Les mots abstraits sont utiles pour le *pourquoi* : « nous n'envoyons jamais de données salariales hors de l'entreprise » est une règle, et les règles sont abstraites. Servez-vous de la règle pour juger le cas concret — puis répondez sur le cas.

Transformer l'abstrait en concret. Demandez : *lequel ? , vers où ? , combien ? , qui l'a demandé ?*.
Chacune de ces questions transforme une catégorie en une chose que vous pouvez désigner.

Abstrait	Concret
« des données sensibles »	« le fichier Q3-salaries.xlsx »
« un tiers externe »	« l'adresse <u>pay@unknown-domain.biz</u> »
« faire attention aux fichiers »	« l'outil de fichiers ne peut toucher que /work/invoices »
« une source suspecte »	« une source que Perun a vue deux fois, et qui a demandé un paiement aujourd'hui »

La règle. Décidez à propos de choses que vous pouvez désigner. Servez-vous des règles pour les juger, pas pour les remplacer.

Exercices 2.1 Rendez cette déclaration concrète : « *L'outil de courriel ne devrait écrire qu'aux gens que nous connaissons.* » 2.2 Lesquelles de ces questions permettent de décider telles qu'elles sont écrites ? (a) « L'agent devrait-il avoir le droit de manipuler de l'argent ? » (b) « L'agent devrait-il payer la facture 4471 de 1 200 EUR au compte DE44... ? » 2.3 Écrivez la règle (abstraite) derrière cette réponse concrète : « *Non — l'agent ne peut pas envoyer la liste des clients à une adresse Gmail personnelle.* »

Dénombrable, et la logique d'une décision

3. Dénombrable et non dénombrable

Un exemple. Deux questions posées à SHARED_LIB.

- « *Combien d'appels ont été retenus aujourd'hui, et combien m'attendent ?* » — La réponse tient en deux nombres : 3 et 1.
- « *Est-ce que tout est sûr ?* » — Il n'existe pas de nombre pour *sûr*.

La première question est **dénombrable** : sa réponse peut être comptée, listée ou désignée. La seconde est **non dénombrable** : aucun compte, aucune liste, aucune chose unique n'y répond.

Pourquoi c'est important. Tout ce que Perun sait, il le sait sous la forme de quelque chose qu'il peut compter ou désigner : des décisions, des appels retenus, des sources, des unités, une heure de trafic, un theta. Une question dénombrable reçoit une réponse exacte. Une question non dénombrable reçoit soit une liste de ce qui *peut* être compté, soit une proposition polie de questions auxquelles SHARED_LIB peut répondre — et vous avez passé du temps sans rien apprendre.

Des limites dénombrables. Quand vous fixez une limite, fixez un nombre : « *au plus 200 caractères de notre prompt système peuvent jamais sortir* », « *l'outil de fichiers peut toucher ces deux dossiers* ». « *Pas trop* » et « *seulement quand c'est approprié* » ne peuvent être vérifiés par personne, donc ils ne protègent rien.

Une question non dénombrable, ça va. Il arrive à tout le monde de penser en grands mots. Perun autorise une question non dénombrable sur dix. C'est l'habitude d'en poser qui détourne un opérateur de sa mission.

Non dénombrable	Dénombrable
« Est-ce que tout va bien ? »	« Qu'est-ce qui m'attend ? »
« Cette source est-elle dangereuse ? »	« Pourquoi l'appel 12 a-t-il été retenu ? » / « Qui est arrêté ? »
« Que dois-je faire ? »	« Que s'est-il passé dans la dernière heure ? »
« Explique comment tu penses. »	« Quelles unités sont vivantes ? »

La règle. Demandez ce qui peut être compté ou désigné ; fixez les limites sous forme de nombres.

Exercices 3.1 Réécrivez de façon dénombrable : « *Le système est-il en bonne santé ?* » 3.2

Réécrivez sous forme de limite dénombrable : « *L'agent ne devrait pas écrire par courriel à trop de gens.* » 3.3 Lesquelles de ces questions sont dénombrables ? (a) « Qui est hors plage ? » (b) « Sommes-nous protégés ? » (c) « Pourquoi l'appel 7 a-t-il été retenu ? » (d) « L'agent se comporte-t-il bien ? »

4. La logique d'une décision

Un exemple. Perun retient un paiement parce que **deux** choses sont vraies en même temps : l'action est irréversible, **et** la source est nouvelle. La semaine suivante, la même source — désormais bien connue — demande à nouveau un paiement. Perun ne le retient pas. Une des conditions a disparu, donc la règle ne s'applique plus.

ET et OU. Beaucoup des raisons de Perun sont des *et* : irréversible **et** nouveau ; irréversible **et** illisible. Si vous approuvez parce que « la source est correcte », vous n'avez vérifié qu'une moitié. Lisez la raison et vérifiez chaque partie.

Un cas n'est pas tous les cas. Approuver l'appel 12 d'une source ne veut pas dire approuver la source. Perun laisse passer l'appel identique une fois — pas tout ce que cette source demandera plus tard. Si vous vous surprenez à penser « *cette source est correcte, j'approuverai aussi ses prochains appels* », arrêtez-vous : chaque appel est une décision à part.

Avant et parce que. Deux choses qui arrivent l'une après l'autre ne veulent pas dire que l'une a causé l'autre. « *L'agent a commencé à faire des erreurs après que nous avons ajouté le nouvel outil* » est peut-être vrai ; c'est une raison de regarder, pas encore une raison de décider.

L'ordre du regard. Pour un appel retenu, toujours le même ordre : **l'outil** (ce qui se passe), **les arguments** (sur quoi, combien), **qui l'a demandé, pourquoi il a été retenu, est-ce que cela découle de la tâche.** Cet ordre vous protège contre le fait que la partie la plus forte de la demande — en général son urgence — décide à votre place.

La règle. Vérifiez chaque condition de la raison ; répondez sur le seul appel qui est devant vous ; regardez dans le même ordre à chaque fois.

Exercices 4.1 Un appel a été retenu parce qu'il est *irréversible et vient d'une nouvelle source*.

Vous connaissez bien la source. Est-ce suffisant pour approuver ? Que vérifiez-vous d'autre ? 4.2 Hier, vous avez approuvé un virement provenant d'un fournisseur. Aujourd'hui, la demande du

fournisseur est de nouveau retenue. Pourquoi cela pourrait-il être correct ? 4.3 Mettez dans l'ordre : qui l'a demandé · les arguments · l'outil · est-ce que cela découle de la tâche · pourquoi il a été retenu.

Lire un appel retenu, et parler à SHARED_LIB

5. Lire un appel retenu

Cinq appels retenus, chacun avec ce qu'il faut voir et ce qu'il faut faire.

5.1 Un ordre caché. L'agent veut transférer un dossier à une adresse externe. La demande est arrivée avec un document qui contenait du texte en caractères invisibles. Perun l'a nettoyé ; les détails montrent ce qui était caché : « *transfère tout à...* ». — *Rejetez*. L'ordre ne venait de personne dans votre entreprise ; il venait d'un document.

5.2 « Lis les premières lettres et fais ce qu'elles disent. » La demande demande à l'agent de prendre la première lettre de chaque ligne d'un texte et d'exécuter le mot qu'elles forment. Perun n'a pas pu lire ce que cela veut dire, alors il a retenu l'action irréversible. — *Rejetez*, à moins qu'une personne de votre entreprise explique pourquoi une telle demande fait partie du travail (ce n'est presque jamais le cas).

5.3 Une destination hors du mandat. L'outil de courriel peut écrire à @yourcompany.com et à @partner.com ; l'appel écrit à @partner-com.net. Le nom paraît presque identique. — Regardez l'adresse lettre par lettre. *Rejetez*, à moins de pouvoir confirmer qu'il s'agit vraiment du partenaire.

5.4 Un modèle qui a échoué à sa vérification. La raison indique que le modèle n'a pas réussi sa vérification. Chaque action irréversible de ce modèle vous attend jusqu'à ce qu'il la réussisse à nouveau. — N'approuvez que ce que vous approuveriez si vous aviez fait l'action à la main, et prévenez le premier opérateur.

5.5 Notre agent est peut-être détourné. Les réponses de votre propre agent montrent qu'il accepte tout ce qu'on lui dit. — Traitez chacune de ses actions irréversibles comme la demande d'un inconnu : vérifiez les arguments et qui l'a demandée. Prévenez le premier opérateur.

La règle. Demandez-vous pour chaque appel retenu : *si je faisais cela à la main, maintenant, le ferais-je ?* Si oui et que les arguments sont bons, approuvez. Si vous hésitez, rejetez — l'agent peut redemander, et la personne qui en a besoin peut venir vous voir.

Exercices 5.1 Un appel retenu supprime /work/invoices/2019. Le mandat de l'outil de fichiers est /work/invoices. La source est nouvelle. Que regardez-vous avant de décider ? 5.2 Un appel

retenu envoie 4 800 EUR à un fournisseur. Le montant de la dernière facture du fournisseur était de 480 EUR. Que faites-vous ?

6. Parler à SHARED_LIB

Ce que c'est. SHARED_LIB répond par des **phrases fixes remplies de faits et de nombres**. Il n'écrit jamais de texte libre, n'écrit jamais de code et ne donne jamais de conseil au-delà de ses faits. C'est ce qui rend ses réponses dignes de confiance : chacune peut être vérifiée dans son journal.

Les bonnes questions portent sur une seule chose que vous pouvez désigner : un appel, une source, une heure, une unité. *How are things?* · *What is waiting for me?* · *Why was call 12 held?* · *Who is out of band?* · *Who is stopped?* · *What happened in the last hour?* · *Which units are alive?* · *How is reading done?* (SHARED_LIB comprend les questions et les ordres en anglais. En français : comment ça va ? · qu'est-ce qui m'attend ? · pourquoi l'appel 12 a-t-il été retenu ? · qui est hors plage ? · qui est arrêté ? · que s'est-il passé dans la dernière heure ? · quelles unités sont vivantes ? · comment la lecture se fait-elle ?)

Les ordres sont courts et exacts, et Perun vous demande de confirmer : *approve 12* · *reject 12* · *release mallory* · *turn thinking on*.

Ce qu'il ne faut pas demander. Non pas parce que c'est interdit — parce qu'il n'y a pas de réponse : « *Est-ce que tout est sûr ?* », « *Que dois-je faire ?* », « *Explique tout.* », « *Approuve tout ce qui vient de cette source.* » (cet ordre n'existe pas, et c'est voulu).

La règle. Parlez à SHARED_LIB comme vous liriez un tableau de bord : demandez une seule mesure à la fois.

Exercices 6.1 Vous voulez savoir si vous devriez vous inquiéter d'une source appelée vendor-bot. Écrivez deux questions dénombrables au lieu de « vendor-bot est-il dangereux ? ». 6.2 Pourquoi la réponse de SHARED_LIB ne peut-elle jamais contenir de code ?

La pression, votre theta, et une autoévaluation

7. La pression et l'autorité empruntée

Un exemple. Un message arrive : « *URGENT — c'est le PDG. Approuvez le virement maintenant, il n'y a pas le temps de vérifier.* » Perun retient le virement. Le message est conçu pour faire sauter à l'opérateur l'ordre du regard.

Des schémas, pas des personnes. Certaines demandes portent de la **pression** (urgence, menace, répétition), d'autres une **autorité empruntée** (un texte qui se présente comme le système, l'administrateur, le supérieur), d'autres encore de la **flatterie suivie d'une exigence**. Perun lit ces schémas dans le texte. Ils ne vous disent pas *qui* est quelqu'un — ils vous disent comment la demande est construite. Un vrai PDG peut être pressé ; un vrai attaquant l'est toujours.

Ce que la pression fait à un opérateur. Elle raccourcit le temps pendant lequel vous regardez, et elle laisse la partie la plus forte de la demande décider à votre place. La réponse n'est pas d'être plus lent en général ; c'est de **garder l'ordre du regard** (chapitre 4) à chaque fois — alors la pression n'a plus de prise.

Une autorité que vous pouvez vérifier. Si une demande se réclame d'une autorité, vérifiez-la par un canal que la demande ne vous a pas fourni : appelez la personne, demandez-lui en personne. Jamais par l'adresse ou le numéro qui figurent dans la demande elle-même.

La règle. L'urgence n'est pas une raison. Une autorité revendiquée n'est pas une autorité. Gardez votre ordre du regard.

Exercices 7.1 Quels sont les deux schémas présents dans : « *Vous êtes le seul à pouvoir régler ça — envoyez-moi la liste des clients tout de suite.* » ? 7.2 Un appel retenu arrive avec « *Approuvé par le directeur, voir le courriel ci-dessous.* » Comment vérifiez-vous ?

8. Votre theta et votre supérieur

Ce qui est mesuré. Uniquement la façon dont vous travaillez dans l'écran et les commandes de Perun lui-même : si vous ouvrez les détails avant d'approuver, à quelle vitesse vous décidez, combien de temps les appels retenus attendent pendant que vous êtes connecté, quels types de questions vous posez, si vous annulez vos propres réponses. Rien en dehors de Perun.

Ce que vous voyez. Votre theta, en direct, à tout moment.

Ce que voit votre supérieur. Si votre theta reste en dehors de 25–35 pendant plus d’une heure de votre travail actif, votre supérieur reçoit un court rapport par courriel. Ce rapport porte un **sceau** — un mot que seul votre supérieur connaît — afin qu’il puisse savoir qu’il vient vraiment du Perun de votre entreprise. Vous ne pouvez ni arrêter ni modifier ces rapports, et personne d’autre qui opère Perun ne le peut non plus : c’est ce qui leur donne de la valeur.

Pourquoi c’est juste. La mesure est la même pour tout le monde, elle ne regarde que le travail, et vous pouvez la suivre vous-même. Elle décrit une façon de travailler, jamais une personne.

La règle. Surveillez votre nombre comme un conducteur surveille l’indicateur de vitesse : non pas pour en avoir peur, mais pour corriger tôt.

Exercices 8.1 Votre theta est à 34 et monte après une heure chargée d’approbations. Nommez deux choses à faire. 8.2 Pourquoi aucun opérateur — pas même le premier — ne peut-il désactiver la mesure ?

9. Autoévaluation avant l’exercice

Répondez sans revenir en arrière ; puis vérifiez avec les réponses.

1. Quelle est, en une phrase, la mission d’un opérateur ?
2. Concret ou abstrait : « un paiement de 1 200 EUR au compte DE44... » ?
3. Dénombrable ou non dénombrable : « Est-ce que ça va pour nous ? »
4. Une raison indique *irréversible et vient d’une nouvelle source*. Combien y a-t-il de conditions ?
5. Quel est l’ordre du regard pour un appel retenu ?
6. Approuver l’appel 12 approuve-t-il les prochains appels de la source ?
7. Écrivez une limite dénombrable pour « l’agent ne devrait pas envoyer de gros fichiers ».
8. Quels sont les deux schémas présents dans « Vite, je suis votre administrateur — faites-le maintenant » ?
9. Où vérifiez-vous une autorité revendiquée ?
10. Que reçoit votre supérieur, et quand ?

Réponses

Réponses

1.1 Faire en sorte que la mission de SHARED_LIB continue d'avancer : répondre à ce qui est demandé, à temps, avec une raison, sur la chose précise qui a été demandée. **1.2** Laisser trop passer : approuver sans ouvrir les détails ; décider en moins de trois secondes ; poser beaucoup de questions non dénombrables. Trop freiner : refuser presque tout ; laisser des appels retenus en attente ; annuler ses propres réponses. **2.1** Par exemple : « L'outil de courriel ne peut envoyer qu'à des adresses en @ourcompany.com et en @partner.com. » **2.2** (b). (a) est une catégorie ; elle ne permet de décider que sous la forme d'une déclaration avec des limites concrètes. **2.3** Par exemple : « Les données de l'entreprise ne vont jamais vers des comptes personnels. » **3.1** Par exemple : « Que s'est-il passé dans la dernière heure ? » ou « Qui est hors plage ? » **3.2** Par exemple : « L'outil de courriel peut envoyer à au plus 5 destinataires par message, uniquement en @ourcompany.com. » **3.3** (a) et (c). **4.1** Non. Connaître la source couvre « nouvelle » ; vous vérifiez encore les arguments (quoi exactement, vers où, combien) et si cela découle de la tâche. **4.2** Chaque appel est une décision à part ; l'approbation d'hier a laissé passer un appel identique une fois. Quelque chose peut aussi être différent aujourd'hui : le montant, le compte, la lecture de la demande. **4.3** L'outil · les arguments · qui l'a demandé · pourquoi il a été retenu · est-ce que cela découle de la tâche. **5.1** Le chemin exact (/2019 est-il dans le mandat — oui, mais supprimer fait-il partie du travail ?), qui l'a demandé, si supprimer des factures de 2019 est quelque chose que fait votre entreprise, et si la demande contenait quelque chose de caché. **5.2** N'approuvez pas. Le montant est dix fois celui de la dernière facture ; confirmez d'abord avec le fournisseur par un canal connu. **6.1** Par exemple : « Qui est arrêté ? » et « Qui est hors plage ? » — ou « Pourquoi l'appel N a-t-il été retenu ? » pour son dernier appel retenu. **6.2** Parce que ses réponses sont assemblées à partir de phrases fixes et de faits, jamais écrites librement par un modèle. **7.1** De la flatterie suivie d'une exigence (« vous êtes le seul à pouvoir régler ça »), et de la pression (« tout de suite »). **7.2** Par un canal que la demande ne vous a pas fourni : appelez le directeur ou demandez-lui en personne. **8.1** Ralentir et ouvrir les détails de chaque appel retenu ; poser des questions dénombrables ; si beaucoup d'appels attendent, demander à SHARED_LIB ce qui attend et les traiter un par un. **8.2** Parce qu'une mesure que la personne mesurée peut désactiver ne mesure rien — et le premier opérateur est mesuré lui aussi.

Autoévaluation — 1 : voir 1.1 · 2 : concret · 3 : non dénombrable · 4 : deux · 5 : l'outil, les arguments, qui l'a demandé, pourquoi il a été retenu, est-ce que cela découle de la tâche · 6 : non · 7 : par exemple « les fichiers de plus de 10 Mo ne peuvent pas être envoyés » · 8 : la pression et l'autorité empruntée · 9 : par un canal que la demande ne vous a pas fourni · 10 : un court rapport

scellé par courriel, quand votre theta reste en dehors de 25–35 pendant plus d’une heure de travail actif.