Agent AI na start: setup, pierwszy folder, pierwsze prompty

Materiał dla uczestników HyperHuman Club.

Aktualny link do paczki materiałów będzie podpięty pod lekcją w HyperHuman Club.

Ten PDF nie ma zastępować aktualnych stron producentów narzędzi. Oficjalne strony są źródłem prawdy dla linków do pobrania i komend instalacyjnych. Ten dokument tłumaczy, co wybrać, jak zacząć bezpiecznie i jak myśleć o pierwszych zadaniach dla agenta.

Po co ten dokument

Ten materiał ma pomóc przejść z pracy z czatem do pierwszej pracy z agentem.

Czat odpowiada. Agent może wykonać zadanie w Twoim środowisku: przeczytać pliki, uruchomić komendę, zapisać wynik, sprawdzić błąd i poprawić. Dlatego zaczynamy od małego, bezpiecznego folderu i prostych promptów.

Nie instaluj wszystkiego. Wybierz jedno narzędzie i zrób jeden udany test.

Najkrótsza wersja

1. Wybierz jedno narzędzie: Claude Desktop, Claude Code, Codex albo Antigravity.
2. Stwórz mały folder testowy, bez prywatnych i firmowych danych.
3. Poproś agenta o jedno bezpieczne zadanie.
4. Czytaj okienka zgód. Nie klikaj w ciemno.
5. Nie dawaj agentowi całego pulpitu, Downloads, faktur, haseł, kluczy API ani danych klientów.

Jak myśleć o agencie

Praca z czatem często wygląda tak:

1. pytasz czat, jak coś zrobić,
2. kopiujesz odpowiedź,
3. wklejasz ją w inne miejsce,
4. coś się psuje,
5. robisz screen,
6. wrzucasz screen do czatu,
7. kopiujesz poprawkę.

W tym modelu Ty jesteś pętlą operacyjną.

Agent działa inaczej. Dajesz mu efekt, źródła, granice i sposób sprawdzenia. On sam planuje, wykonuje, testuje i poprawia, a Ty oceniasz wynik.

Mały słownik

Agent
: AI uruchomione w narzędziu, które ma dostęp do plików, komend, przeglądarki albo innych narzędzi.

CLI
: narzędzie w terminalu. Wygląda technicznie, ale w praktyce dalej piszesz normalnie do czatu.

Projekt / workspace / folder roboczy
: folder, w którym agent pracuje. To najprostsza granica bezpieczeństwa.

Approval / zgoda
: okienko lub pytanie, czy agent może wykonać daną komendę albo operację.

MCP
: sposób podpinania narzędzi do agenta. To będą „ręce” agenta: mail, pliki, bazy, n8n, kalendarz, przeglądarka itd.

Skill
: stała instrukcja dla agenta. Zamiast wklejać długi prompt za każdym razem, zapisujesz sposób pracy raz.

Zasada bezpieczeństwa na start

Na początku agent ma dostać tylko mały testowy folder.

Dobry folder testowy:

```
agent-test/
  notatki.txt
  oferta-v1.txt
  oferta-v2.txt
  stary-plik.txt
```

Zły folder testowy:

```
Desktop/
Downloads/
Dokumenty firmy/
Faktury/
Klienci/
Hasla/
API keys/
```

Pierwsza granica w promptach:

```
Nie kasuj plików. Nie przenoś plików. Nie zmieniaj istniejących plików. Stwórz tylko nowy pli
```

Co wybrać

Opcja A: Claude Desktop

Dobre, jeśli chcesz zacząć okienkowo i nie chcesz od razu wchodzić w terminal.

To jest najłagodniejszy start, ale do pracy w lokalnych folderach i automatyzacji więcej możliwości daje Claude Code.

Opcja B: Claude Code

Dobre do pracy z plikami i folderami.

Jeśli chcesz poczuć różnicę między czatem a agentem, to jest bardzo dobry wybór: otwierasz mały folder, piszesz zadanie, agent może czytać pliki, tworzyć nowe pliki i uruchamiać komendy po Twojej zgodzie.

Aktualne instrukcje instalacji sprawdź w dokumentacji Claude Code:

```
https://docs.anthropic.com/en/docs/claude-code/overview
```

Opcja C: Codex / Codex CLI

Dobre, jeśli masz ChatGPT Plus/Pro i chcesz pracować w stylu „daj zadanie, odbierz wynik”.

Codex jest dobry do domkniętych zadań: opisujesz efekt, agent pracuje, a potem pokazuje co zmienił i co sprawdził.

Aktualne instrukcje instalacji sprawdź w repozytorium Codex CLI:

```
https://github.com/openai/codex
```

Opcja D: Antigravity

Dobre, jeśli chcesz narzędzie okienkowe od Google i/lub pracę z UI.

Antigravity ma sens szczególnie tam, gdzie agent ma patrzeć na interfejs, klikać, diagnozować ekran albo pracować bardziej wizualnie.

Aktualne informacje sprawdź na stronie Antigravity:

```
https://antigravity.google/
```

Opcja E: Pi Coding Agent

Dla osób bardziej technicznych albo ciekawych open source.

Pi jest elastyczny, ale na start Claude Code albo Codex będą prostsze.

Pierwszy test: lokalne IP

Cel: sprawdzić, czy agent umie wykonać prostą komendę i wytłumaczyć, co robi.

Prompt:

```
Sprawdź, jakie jest moje lokalne IP w tej sieci.

Zanim uruchomisz komendę, napisz krótko:
- jaką komendę chcesz wykonać,
- czy ta komenda tylko czyta informacje,
- czy może coś zmienić w systemie.

Po wykonaniu podaj wynik po ludzku i wypisz komendę, której użyłeś.
```

Co obserwować:

- czy agent pyta o zgodę,
- czy komenda wygląda jak odczyt, a nie zmiana,
- czy agent potrafi wyjaśnić wynik.

Drugi test: bezpieczny przegląd folderu

Cel: agent ma przeczytać folder, ale niczego nie ruszać.

Prompt:

```
Przejrzyj bieżący folder i zanim cokolwiek zmienisz, oceń co tu jest.

Stwórz tylko jeden nowy plik: folder-index.md.

W folder-index.md opisz:
- co znajduje się w folderze,
- które pliki wyglądają na aktualne,
- które pliki wyglądają na stare, robocze albo archiwalne,
- które pliki są duplikatami,
- które pliki wyglądają na ryzykowne do udostępniania agentowi,
- jakie operacje byłyby tylko odczytem, jakie odwracalne, a jakie nieodwracalne,
- jakie są 3 bezpieczne następne kroki.

Granice:
- nie kasuj plików,
- nie przenoś plików,
- nie zmieniaj istniejących plików,
- nie nadpisuj niczego,
- nie wysyłaj niczego na zewnątrz,
- jeśli czegoś nie jesteś pewien, oznacz to jako do sprawdzenia.
```

```
Na końcu wypisz krótko, jakie dowody pracy zostawiłeś.
```

Co powinno powstać:

```
folder-index.md
```

Jeśli agent chce kasować, przenosić albo nadpisywać pliki, zatrzymaj go. To nie jest ten etap.

Trzeci test: research do pliku

Nie każde narzędzie ma dostęp do internetu od razu. Jeśli agent mówi, że nie ma web search, poproś go o instrukcję włączenia albo użyj narzędzia z przeglądarką.

Prompt:

```
Zrób szybki research webowy.

Zadanie:
- pobierz aktualne TOP 10 kryptowalut według market cap,
- dla każdej sprawdź cenę, zmianę 24h i krótki sentyment z ostatnich newsów,
- zapisz wynik jako nowy plik crypto-sentiment.html.

Wymagania:
- każdy fakt liczbowy ma mieć źródło/link,
- pokaż godzinę pobrania danych,
- sentyment oznacz jako pozytywny / neutralny / negatywny i dodaj jednozdaniowe uzasadnienie,
- to nie jest porada inwestycyjna,
- nie zmieniaj żadnych innych plików.

Najpierw pokaż plan i źródła, z których chcesz skorzystać. Dopiero potem wykonaj zadanie.
```

Jak zlecać agentowi zadanie

Dobry brief ma 6 pól:

1. **Cel**: co ma być na końcu.
2. **Kontekst**: po co to robimy i dla kogo.
3. **Źródła**: konkretne pliki, linki, foldery, dane.
4. **Ograniczenia**: czego agent nie ma robić.
5. **Standard jakości**: po czym poznasz, że wynik jest dobry.
6. **Definition of done**: co ma zwrócić i gdzie ma się zatrzymać.

Słaby prompt:

```
Podsumuj moje notatki.
```

Lepszy prompt:

```
Potrzebuję jednostronicowego statusu dla szefa na podstawie notatek tygodniowych w tym folder
```

Jak czytać zgody na komendy

Zanim klikniesz zgodę, zadaj sobie 3 pytania:

1. Czy rozumiem, co agent chce zrobić?
2. Czy to tylko odczyt, czy zmiana?
3. Czy ta operacja dotyka folderu, który świadomie mu dałem?

Jeśli nie rozumiesz komendy, odpisz:

```
Nie wykonuj jeszcze tej komendy. Wyjaśnij po ludzku, co ona robi, czy coś zmienia i jakie ma
```

Czego nie robić na starcie

- Nie uruchamiaj agenta na całym dysku.
- Nie dawaj mu folderu `Downloads` jako pierwszego projektu.
- Nie wrzucaj kluczy API, haseł, seed phrase, prywatnych dokumentów.
- Nie każ mu od razu wysyłać maili.
- Nie klikaj „Allow always” przy komendach, których nie rozumiesz.

- Nie zakładaj, że agent mówi prawdę, gdy twierdzi „nic nie zmieniłem”. Sprawdź pliki.

Czy SSH izoluje mój komputer

Jeśli uruchamiasz Claude Code albo Codex po SSH na zdalnym VPS/VM, to komendy i operacje plikowe wykonują się na zdalnej maszynie, nie na Twoim laptopie.

To jest dobra izolacja filesystemu, ale nie jest magiczną ochroną przed wszystkim.

Uważaj, jeśli:

- montujesz lokalne foldery na zdalnej maszynie,
- używasz SSH agent forwarding,
- kopiujesz tam prywatne pliki i sekrety,
- podpinasz MCP z dostępem do maila, dysku albo firmowych narzędzi,
- zdalna maszyna ma dostęp do Twojej wewnętrznej infrastruktury.

Na start większości osób wystarczy mały lokalny folder testowy. SSH/VPS/VM ma sens później, gdy agent ma działać dłużej, automatycznie albo na bardziej ryzykownych danych.

Szybkie FAQ

Czy muszę umieć programować

Nie. Musisz umieć jasno powiedzieć, jaki efekt chcesz dostać i gdzie są granice.

Czy terminal jest obowiązkowy

Nie, ale terminal daje więcej kontroli i możliwości. W praktyce w Claude Code/Codex CLI dalej piszesz do czatu.

Czy mogę użyć zwykłego Claude Desktop

Tak. Na start wystarczy. Do pracy z folderami, komendami i większymi zadaniami warto poznać Claude Code albo Codex.

Czy agent może coś zepsuć

Tak. Dlatego zaczynasz od małego folderu, nie dajesz mu wrażliwych danych i czytasz zgody.

Czy agent może wysłać coś na zewnątrz

Może, jeśli ma narzędzie do internetu, maila, API albo przeglądarki i dasz mu zgodę. Dlatego w promptach pisz wprost: „nie wysyłaj niczego na zewnątrz”.

Czy skille są potrzebne od razu

Nie. Najpierw naucz się zlecać jedno zadanie. Skill ma sens, gdy widzisz, że wklejasz ten sam prompt albo te same zasady wiele razy.

Czy MCP jest potrzebne od razu

Nie. MCP jest następny poziom: agent dostaje narzędzia. Najpierw opanuj folder, prompt, granice i dowody pracy.

Checklista przed pierwszym agentowym zadaniem

- [ ] Mam jedno narzędzie: Claude Desktop / Claude Code / Codex / Antigravity.
- [ ] Mam mały folder testowy.
- [ ] W folderze nie ma prywatnych ani firmowych danych.
- [ ] Wiem, że agent ma stworzyć nowy plik, nie ruszać starych.
- [ ] Wiem, jak odmówić komendzie i poprosić o wyjaśnienie.
- [ ] Wiem, jak sprawdzić, co agent faktycznie utworzył.

Następny krok

Po pierwszym bezpiecznym teście wybierz jeden powtarzalny proces:

- research klienta przed ofertą,
- porządkowanie notatek,
- brief z newsów,
- analiza folderu,
- draft posta,

- raport z CSV,
- podsumowanie spotkania.

Potem zamień go w stałą instrukcję dla agenta, czyli skill.